UNIVERSITY OF RIJEKA
FACULTY OF ENGINEERING

Mateja Napravnik

# FOUNDATION MODELS FOR TRANSFER LEARNING TRAINED ON THE RADIOLOGYNET MEDICAL DATASET

DOCTORAL THESIS

Rijeka 2025.

UNIVERSITY OF RIJEKA
FACULTY OF ENGINEERING

Mateja Napravnik

# FOUNDATION MODELS FOR TRANSFER LEARNING TRAINED ON THE RADIOLOGYNET MEDICAL DATASET

DOCTORAL THESIS

Supervisor: Prof. Ivan Štajduhar, PhD

Rijeka 2025.

SVEUČILIŠTE U RIJECI
TEHNIČKI FAKULTET

Mateja Napravnik

# TEMELJNI MODELI STROJNOGA UČENJA ZA PRIJENOS ZNANJA OBUČAVANI NA SKUPU MEDICINSKIH PODATAKA RADIOLOGYNET

DOKTORSKI RAD

Mentor: prof. dr. sc. Ivan Štajduhar

Rijeka 2025.

Doctoral thesis supervisor: Prof. Ivan Štajduhar, PhD (Faculty of Engineering, University of Rijeka)

The doctoral thesis was defended on _____ at the University of Rijeka, Faculty of Engineering, Croatia, in front of the following Evaluation Committee:

1. Prof. Ivo Ipšić, PhD (Faculty of Engineering, University of Rijeka) - Committee Chair
2. Prof. Sandi Ljubić, PhD (Faculty of Engineering, University of Rijeka)
3. Assist. Prof. Žiga Emeršič, PhD (Faculty of Computer and Information Science, University of Ljubljana, Slovenia)

This work has been fully supported by the Croatian Science Foundation/Hrvatska zaklada za znanost [grant number IP-2020-02-3770, DOK-2021-02-6744] (*Machine learning for knowledge transfer in medical radiology / Strojno učenje za prijenos znanja u medicinskoj radiologiji*).

This research also used resources from the Center for Advanced Computing and Modelling, University of Rijeka.

# ACKNOWLEDGEMENTS

An honest thank you to my supervisor and mentor, prof. Ivan Štajduhar, PhD, for his patience, guidance, advice, and supportive shoulder taps. I am also deeply grateful to my friend and colleague, Franko Hržić, PhD, for his invaluable contributions and unwavering support throughout this journey.

Thank you to my friends for being there when I needed it. I would also like to thank my co-workers for the great moments we shared – your camaraderie and good humour made working at the Faculty of Engineering a true pleasure.

This would not be possible without the help and support of my family – my parents Marija and Željko, and my dearest Stjepan (*Bado*). There are no words I can use to express my gratitude.

*To my sister.*

*Mojoj divnoj, nezaboravljenoj, nezamjenjivoj sestri.*

1992 – †2014

# ABSTRACT

The adoption of deep learning techniques in medical imaging has the potential to improve diagnostic accuracy and speed up clinical decision-making. However, the development of such techniques is slowed down by the scarcity of annotated datasets, as manual labelling of medical data is time-consuming, costly, and expert-dependent. For this reason, transfer learning has been widely adopted as a solution: a model is first pretrained on a large dataset, and then fine-tuned on downstream tasks (which are often data-scarce). However, publicly available large-scale medical datasets often focus narrowly on specific imaging modalities or anatomical regions (e.g. chest X-rays), thereby restricting their usefulness in constructing general-purpose models for transfer learning. The reliance on natural image datasets (e.g. ImageNet) for pretraining has shown mixed results in medical transfer learning, which underscores the need for large and diverse meaningful medical datasets that can be used in the development of pretrained models.

This thesis addresses the lack of domain-relevant annotated data by introducing an unsupervised framework for labelling medical imaging datasets using a combination of Digital Imaging and Communications in Medicine (DICOM) images, structured metadata, and narrative diagnoses. The pipeline was applied to a large-scale multimodal medical dataset, RadiologyNET, with feature extraction and clustering techniques used to group images into semantically meaningful categories without relying on manual annotation. These pseudo-labels were then used to pretrain several widely used convolutional neural network architectures, including ResNet, EfficientNet, DenseNet, MobileNet, Inception and VGG.

The pretrained models were evaluated on a wide range of downstream tasks (classification, regression, and segmentation) across multiple publicly available medical imaging datasets. Comparative analyses were conducted against both ImageNet-pretrained

models and models trained from randomly initialised weights. The findings show that RadiologyNET-pretrained models are effective when training resources are limited (i.e. reduced training data and training time), however, they did not consistently outperform ImageNet in normal training conditions. ImageNet-pretrained models achieve strong performance when fine-tuned, but the overall benefits of transfer learning (regardless of source) decrease as the amount of available training data increases, confirming that the impact of pretraining becomes less prominent in problems with sufficient data.

**Keywords: Transfer Learning, DICOM, Foundation Models, Pretraining, Medical Image Analysis, Machine Learning**

# PROŠIRENI SAŽETAK

Primjena tehnika dubokog učenja u medicini može poboljšati dijagnostiku i ubrzati donošenje kliničkih odluka. Međutim, razvoj takvih modela je usporen zbog nedostatka označenih podatkovnih skupova, budući da je ručno označavanje medicinskih podataka skupo, vremenski zahtjevno i mogu ga odraditi samo medicinski stručnjaci (liječnici). Učenje s prijenosom znanja je postala široko-prihvaćena metoda za ublažavanje efekta nedostatka podataka, a u tom procesu se prvo model predtrenira na velikom podatkovnom skupu, nakon čega se njegovi parametri koriste za treniranje na ciljnim podacima (kojih često nema dovoljno). Međutim, javno dostupni medicinski skupovi podataka često su usmjereni na mali raspon modaliteta snimanja ili anatomskih regija, čime se ograničava njihova korisnost u razvoju široko-namjenskih predtreniranih modela u medicini. Oslanjanje na velike skupove podataka prirodnih slika (primjerice ImageNet) za predtreniranje je pokazalo neujednačene rezultate u kontekstu prijenosa znanja na medicinskim zadacima, što dodatno naglašava potrebu za dovoljno velikim i raznolikim medicinskim skupovima podataka za razvoj temeljnih modela.

Ovaj doktorski rad pokušava riješiti problem nedostatka (dovoljno velikih) označenih medicinskih skupova uvođenjem nenadgledanog okvira za označavanje medicinskih podataka koristeći kombinaciju radioloških slika, pripadajućih metapodataka i tekstualnih dijagnoza. Navedeni se sustav sastoji od različitih tehnika ekstrakcije značajki i grupiranja s ciljem svrstavanja slika u semantički slične kategorije bez potrebe za ručnim označavanjem, te je primijenjen na veliki multimodalni medicinski skup podataka RadiologyNET. Dobivene pseudo-oznake korištene su za predtreniranje nekoliko široko rasprostranjenih arhitektura konvolucijskih neuronskih mreža, uključujući ResNet, EfficientNet, DenseNet, MobileNet, Inception i VGG.

Predtrenirani modeli vrednovani su na različitim zadacima (klasifikacija, regresija i

segmentacija) koristeći više javno dostupnih medicinskih slikovnih skupova podataka. Provedena je analiza s modelima predtreniranim na ImageNetu, kao i s modelima treniranima iz nasumično inicijaliziranih parametara. Rezultati pokazuju da modeli predtrenirani na skupu podataka RadiologyNET postižu dobre rezultate u uvjetima gdje su resursi za treniranje ograničeni (primjerice, smanjeni broj podataka ili vrijeme treniranja). Pri tome, u standardnim (neograničenim) uvjetima modeli RadiologyNET nisu signifikantno nadmašili modele predtrenirane na ImageNetu, nego su pokazali slične performanse. U eksperimentima se pokazalo da se opća korisnost učenja prijenosom znanja (bez obzira na izvor) smanjuje što je više dostupnih podataka za treniranje, potvrđujući da je utjecaj prijenosa znanja manji kod problema s dovoljno podataka.

**Ključne riječi: Učenje prijenosom znanja, DICOM, Temeljni modeli, Predtreniranje, Analiza medicinskih slika, Strojno učenje**

# CONTENTS

# 1. Chapter

# INTRODUCTION

Over the past decade, machine learning and deep learning have seen increasing adoption in medical diagnosis and treatment [1]. Computer-aided diagnosis systems [2] based on neural networks have demonstrated performance on par with or even exceeding that of human experts in certain tasks, such as melanoma detection [3] or COVID-19 classification from chest radiographs [4, 5]. Convolutional neural networks (CNNs) have become a popular choice for medical image classification, segmentation, and regression [6]; and more recently, vision transformers are emerging as an alternative [7, 8, 9]. Despite these advances, the development of robust machine learning models in medical imaging remains limited by the (un)availability of annotated datasets [10]. High-quality annotation requires specialised clinical expertise, making the process time-consuming and costly [11, 12]. Furthermore, different hospital centres may have different terminology or labelling practices, making the process inconsistent across institutions. To mitigate this overall lack of high-quality annotated medical data, there is a consensus among researchers that leveraging transfer learning (TL) is the path forward in machine learning [12].

In TL, a model is first pretrained on large datasets with sufficient amounts of data, and then retrained or fine-tuned on the actual specific dataset of the target task. Fine-tuning is the process of adapting a pretrained model to a new task by continuing the training on the target dataset. This typically involves adjusting the model weights to better align with the characteristics of the new data while retaining the general knowledge acquired during pretraining. This approach is highly effective in medical imaging tasks where domain-specific data is often limited, and it often leads to better model stability [10] (thereby

lessening the impact of annotated data scarcity). These pretrained models, which can be used as starting points in a wide range of downstream tasks, are often referred to as *foundation models*.

ImageNet [13, 14], a large-scale dataset comprising millions of natural images, has become a standard resource for pretraining deep learning models. Its use in medical machine learning remains widespread, largely due to its accessibility and demonstrated performance improvements in a variety of downstream tasks [15, 16]. However, some studies question its suitability for medical applications, with the reason being that the domain shift between natural and medical images (i.e. in semantics and structure) limit the effectiveness of ImageNet-based TL in clinical settings [11, 17, 18]. In response, recent research has seen a rise of medical foundation models such as RadImageNet and BiomedCLIP [19, 20], which offer pretrained models better aligned with the requirements of medical tasks. Nevertheless, ImageNet remains a prevalent starting point in many medical machine learning pipelines [15], with a recent study by Woerner et al. [16] showing that, when ImageNet-pretrained models are fine-tuned, their performance is on-par with many medical foundation models. While numerous studies explore different pretraining strategies, ranging from supervised learning on medical datasets to self-supervised and contrastive approaches [21, 22], the rationale behind selecting a particular dataset or pretraining method is seldom explicitly stated [11, 17]. Although some studies do offer guidance, comprehensive repositories of pretrained medical models across diverse architectures and tasks remain scarce [22, 23].

To this end, this thesis proposes the use of a custom, large-scale medical imaging dataset called RadiologyNET, which contains 2.3 million examinations conducted at the Clinical Hospital Centre Rijeka between 2008 and 2017. RadiologyNET spans multiple imaging modalities (e.g. computed tomography - CT, magnetic resonance - MR, Computed Radiography - CR) and anatomical regions (e.g. head, abdomen), making it a large-scale medical dataset suitable for building domain-specific foundation models.

## 1.1. Related Work

The use of foundation models has become increasingly prevalent in medical machine learning as a means of addressing challenges related to data (un)availability, domain speci-

ficity, and generalisability. Altough TL-based strategies have demonstrated improvements in medical imaging tasks, building such models requires a significant amount of resources: sufficient data and time, as well as meeting high computational demands of modern deep learning.

A central challenge in developing medical foundation models lies in the availability of high-quality annotations. *Supervised learning*, which forms the basis of most clinical diagnostic models, depends on these labels to learn mappings from input data to clinically relevant outputs. Although models pretrained using supervised learning can capture the semantic meaning of each pixel through linking pathologies with patterns found in images, acquiring expert-labelled medical data is a difficult task. To overcome this, alternative pretraining strategies have gained popularity. *Unsupervised learning* leverages the data structure without requiring labels, often using clustering or dimensionality reduction to extract patterns. *Self-supervised* learning creates pseudo-labels from the data itself, and an example of this is *contrastive learning*, which trains models to distinguish between similar and dissimilar data points. These approaches allow foundation models to be pretrained on large unlabelled datasets, which is particularly valuable in the medical domain where annotated data is scarce.

This section is organised into three parts to systematically address the landscape of medical foundation models and the datasets utilised their development. The first subsection, *Overview of available medical foundation models*, presents a review of recent notable models specifically designed for medical applications, describing their training strategies, target domains, and limitations. The second subsection, *Existing annotated medical datasets*, provides a comparative analysis of existing large-scale datasets that have been (or could be) used in model pretraining, discussing their scope, annotation strategies, and modality/anatomical region coverage. The third subsection, *Automated annotation and pseudo-labels*, examines methods whose goal is to overcome the limitations of manual labelling through, for example, unsupervised and self-supervised approaches, including clustering, contrastive learning, and representation learning without explicit annotations. Together, these subsections provide a background on the current state of medical foundation models for TL.

### 1.1.1. Overview of available medical foundation models

**Table 1.1:** Examples of publicly available medical foundation models.

| **Foundation Model** | **Training Data** | **General Purpose / Application** |
|---|---|---|
| BioMedCLIP [19] | Radiology images paired with text reports | Vision-language retrieval, classification, multimodal tasks |
| GatorTron [24] | 90 billion words of clinical text (e.g. from medical literature) | Clinical language processing, concept extraction, question answering |
| TotalSegmentator [25] | Multi-organ CT datasets | Multi-organ segmentation, anatomical structure delineation |
| nnU-Net [26] | Diverse medical image segmentation datasets | Self-configuring segmentation across multiple modalities and organs |
| SAM-Med Family [27] (MedSAM, SAM-Med2D, SAM-Med3D) | Over 1.5M 2D image–mask pairs; 22,000 3D images with 143,000 masks | Promptable 2D/3D segmentation, zero-shot segmentation across modalities |
| Me-LLaMa [28] | Instruction-tuned on medical question–answer datasets (based on LLaMA) | Clinical dialogue generation, medical question answering, decision support |

The development of foundation models has significantly accelerated progress in natural image analysis and natural language processing. In recent years, similar approaches have been extended to the medical domain, where the objective is to pretrain large, generalisable models on diverse datasets, thus improving fine-tuning across a wide range of clinical tasks. Several foundation models specifically tailored for medical imaging and medical text analysis have been built and released, with several examples presented in Table 1.1 and described here.

**BioMedCLIP** is an example of a vision-language foundation model adapted for the medical domain. Inspired by Contrastive Language–Image Pretraining (CLIP) in the natural image space, BioMedCLIP pretrains its vision encoder and text encoder jointly using paired radiology reports and images using contrastive learning. The model has demonstrated strong performance across multiple retrieval and classification tasks, particularly in matching medical images with clinical text [19]. BioMedCLIP represents a key step

towards multimodal understanding in healthcare applications, as it is suitable for various downstream adaptations in vision-language problems in medicine.

In the field of medical language processing, **GatorTron** [24] is a large-scale foundation model trained on over 90 billion words of clinical text. GatorTron leverages transformer-based architectures similar to Bidirectional Encoder Representations from Transformers (BERT) [29] but is adapted specifically for medical applications such as clinical concept extraction, medical question answering, and document classification. By pretraining on domain-specific corpora, GatorTron has achieved state-of-the-art performance across a variety of medical language processing benchmarks, which highlights the value of domain-specific pretraining in text-based clinical tasks. **Me-LLaMa** [28] also represents one such initiative, building on the LLaMA model family through domain-specific instruction tuning. Rather than training from scratch, MeLLama fine-tunes general-purpose large-language models using clinical datasets and medical question–answer pairs, making it suitable for medical dialogue and clinical decision support.

**nnU-Net** [26] has become a widely recognised framework in medical image segmentation. Although it is not a foundation model in the strictest sense, nnU-Net can automatically adjust itself to new segmentation tasks (without manual configuration), which makes it highly generalisable. Its design principles (automatic pre-processing, architecture selection, and hyperparameter tuning) have established strong baselines across many medical imaging challenges, and the success of nnU-Net's task-adaptive model design has influenced subsequent work on building scalable and widely applicable models in medical imaging. This led to the developmnent of **TotalSegmentator** [25], which was built upon the basis of nnU-Net. Trained on extensive multi-organ CT datasets, TotalSegmentator is designed to perform full-body organ segmentation across more than 100 anatomical structures. Unlike traditional segmentation models focused on single-organ tasks, TotalSegmentator demonstrates strong generalisability without retraining or fine-tuning, which makes it a valuable tool for a wide range of clinical and research applications.

Beyond task-specific segmentation frameworks, recent work has also focused on developing fully generalisable segmentation foundation models capable of zero-shot performance across modalities and anatomical targets. To this end, the **SAM-Med family** [27] is another step forward in medical image segmentation. This model family consists of MedSAM, SAM-Med2D, and SAM-Med3D, and its goal is to address the generalisability

across varying imaging modalities and disease types in addition to anatomical structures. MedSAM was trained on over 1.5 million image–mask pairs spanning 10 imaging modalities and more than 30 cancer types, resulting in a model capable of robust and accurate segmentation across diverse clinical contexts. Building on this, SAM-Med3D extends the concept to volumetric (3D) medical data, utilising a fully learnable 3D architecture trained on a dataset comprising 22,000 3D images and 143,000 segmentation masks. Incorporating prompt-based segmentation, SAM-Med3D demonstrates good transferability to unseen anatomical targets and imaging modalities without additional retraining or fine-tuning.

Despite recent progress, many medical foundation models remain constrained to narrow domains (e.g. chest X-rays, brain MRI, CT scans) or specific tasks (e.g. segmentation of anatomical regions). At the same time, existing foundation models are often large (complex), contain millions (or billions) of parameters, and are highly-demanding in terms of hardware, which can limit their practicality, portability and ease-of-use. As a consequence, TL based on *traditional*, widely used network architectures (e.g. ResNet [30], DenseNet [31], EfficientNet [32]) remains popular. Although these *traditional* models may not always achieve state-of-the-art performance (as state-of-the-art performance usually comes with mechanisms specifically tailored to the targeted task), they offer noticeable advantages in simplicity and accessibility [11].

### 1.1.2. Existing annotated medical datasets

Training foundation models for TL in the medical domain requires access to large-scale, diverse medical datasets, which circles back to the issue of medical data scarcity. Furthermore, such datasets are rarely made publicly available due to the difficulty of anonymising medical records. While some large datasets do exist, they tend to be limited in scope, focusing on specific imaging modalities or anatomical regions (e.g. the popular CheXpert dataset includes only chest X-ray images [33]). This lack of coverage restricts the development of general-purpose models applicable across a broader clinical spectrum.

Over the years, several datasets have been released that serve as benchmarks for various tasks such as classification, segmentation, and report generation. However, they often differ widely in terms of modality coverage, annotation granularity, and their suitability

**Table 1.2:** Overview of various medical imaging datasets.

| Dataset | Modality | Size | Annotations | Applications | Features |
|---|---|---|---|---|---|
| DeepLesion [34] | CT | 32,120 CT slices | Lesion bounding boxes | Lesion detection and classification | Diverse lesion types from multiple body regions |
| RadImageNet [20] | CT, MR, ultrasound | 1,350,000 images | 165 pathologies/labels | Pretraining for medical imaging AI | Large-scale dataset for transfer learning |
| CheXpert [33] | X-ray | 224,316 images | 14 common observations (e.g. pneumonia) | Chest disease classification | Uncertainty labels for pathologies |
| ChestX-ray14 [35] | X-ray | 108,948 images | 8 text extracted labels | Chest disease classification | One of the largest publicly available chest X-ray datasets |
| MIMIC-CXR [36] | X-ray | 377,110 images | Radiology reports, Textual diagnoses | Image and text generation | Paired image-text dataset with free-text radiology reports |
| MedPix 2.0 [37] | Various | $\approx$ 59,000 images | 12,000 cases | Multimodal medical education | Designed for teaching and multimodal AI applications |
| MURA [38] | X-ray | 40,561 images | Binary abnormality labels | Musculoskeletal abnormality detection | Focused on upper extremity abnormalities |
| OASIS [39, 40] | MR | $\approx$ 1200 sessions (OASIS-1) $\approx$ 2000 sessions (OASIS-3) A | Brain structure and dementia-related labels | Neuroimaging research | Longitudinal data for Alzheimer's disease research |

for general-purpose model pretraining. Table 1.2 provides an overview of several popular medical imaging datasets.

The **DeepLesion dataset** [34] is a large, publicly available collection of over 32,000 annotated lesions on CT images which covers a wide range of lesion types and anatomical regions. Annotations were extracted from routine radiologist bookmarks, making the dataset highly representative of actual practice. Although this dataset is diverse in terms of anatomical regions, it is limited to a single imaging modality (CT), which restricts its utility in cross-modality model development.

The aforementioned **CheXpert** dataset is a large, publicly available dataset comprising 224,316 chest radiographs from 65,240 patients, collected at Stanford Health Care [33]. It includes both frontal and lateral views, with each image labelled for the presence of 14 observations, such as various lung diseases, using a system that accounts for uncertainty in radiological interpretation. Similarly, **ChestX-ray14** [35] includes over 100,000 frontal-view chest X-rays, labelled with eight disease categories extracted automatically from radiology reports using natural language processing. **MIMIC-CXR** [36] offers over 370,000 chest X-ray images, paired with free-text radiology reports. This pairing makes it well-suited for multimodal tasks such as report generation, cross-modal retrieval, and vision-language pretraining. While these datasets have contributed towards thoracic disease classification and content-based image retrieval, their focus on a single imaging modality and anatomical region limits their applicability for developing general-purpose medical models. Similarly, **MURA** [38] is another example of an X-ray image dataset, but it is focused on musculoskeletal pathologies and contains over 40,000 images labelled as normal or abnormal. It targets classification tasks involving the upper extremities and has been used in studies focusing on wrist, elbow, finger, and shoulder abnormalities. Although it covers a wider range anatomical regions compared to CheXpert or ChestX-ray, it is still limited solely to X-ray images.

**MedPix 2.0** is a freely available biomedical dataset featuring over 12,000 clinical cases and nearly 59,000 medical images. It pairs images with detailed text such as findings and diagnoses, and contains images captured in different radiology imaging modalities (CT and MR imaging). However, it is not as large as some other datasets in terms of raw image volume.

It was specifically designed for neuroimaging research, and newer versions of OASIS

(OASIS-3 [40]) include longitudinal scans, making the dataset well-suited for research studying disease progression (i.e. changes in the brain over time, such as Alzheimer's progression).

Out of the presented datasets, **RadImageNet** is the largest and most suitable for building general-purpose models for TL in the medical domain, which is precisely the purpose it was created for. It includes over 1.3 million images across modalities such as CT, MR, and ultrasound, annotated with 165 pathology labels. It marks a significant effort in terms of building a dataset for TL, as it required a team of 20 expert radiologists to label each of the 1.3 million images. RadImageNet was used to pretrain several CNNs, which were then fine-tuned on downstream medical tasks. Initial results have shown performance gains over ImageNet-pretrained models in several tasks.

### 1.1.3. Automated annotation and pseudo-labels

As demonstrated by RadImageNet [20], annotating over one million images with high-quality pathology labels required the dedicated effort of 20 expert radiologists. This level of manual annotation, while achievable for a limited set of predefined labels, is not scalable for increasingly diverse, multimodal clinical datasets.

RadiologyNET [41] is a large medical dataset collected from the Clinical Hospital Centre Rijeka, which was acquired through standard clinical practice between 2008 and 2017. The entire dataset consists of 2.3 million examinations and nearly 25 million Digital Imaging and Communications in Medicine (DICOM) [42] files. Its total size is approximately 13 terabytes, spanning multiple imaging modalities commonly encountered in clinical practice, including CT, MR, CR, X-ray angiography (XA), nuclear medicine (NM), and radiofluoroscopy (RF), among others. Manual annotation of a dataset of this magnitude would not be feasible, and it would exceed the resources invested in previous datasets such as RadImageNet (as there are 25 million images in the total RadiologyNET dataset, versus the 1.3 million in RadImageNet).

Although DICOM files store both the image data and structured metadata [42], the latter is often manually entered by clinicians and can be inconsistent or error-prone. Errors range from typographical mistakes to inconsistent labelling practices (e.g. one physician may indicate an image depicts a *leg*, while the other may label it as *extrem-*

*ity* or *lower extremity*). As a result, metadata alone is often insufficient for reliable labelling, and raw pixel data typically lacks the semantic context required to differentiate between similar images; for example, *elbow* and *knee* X-ray images often have similar pixel value distributions. As manual annotation of large datasets remains both time-consuming and resource-intensive [12, 20], this has prompted growing interest in label-efficient approaches, particularly self-supervised and unsupervised learning methods. By identifying patterns within the data, these methods can group semantically similar images together, reducing the reliance on manual input. While self-supervised and unsupervised methods may not achieve the same level of precision as manual labelling by experts, they can provide a valuable starting point for annotation tasks and accelerate the annotation process [20]. For example, contrastive learning (which is self-supervised) has been shown to effectively group images based on semantic similarity, later enabling classification using learned feature embeddings [43]. Similarly, Guo et al. demonstrated that autoencoders combined with k-means clustering can be used to extract relevant features and group images, facilitating annotation for subsequent tasks [44].

The choice of feature extraction technique often depends on the data type. For structured tabular data, methods such as principal component analysis (PCA) [45] or autoencoders [46] are widely used for feature extraction. In natural language processing, Bag of Words remains a common approach for converting text into feature vectors [47]. For image data, simple methods like Histogram of Oriented Gradients have been used in the past, though neural networks have risen in popularity in recent years due to their ability to learn complex representations [48].

## 1.2. Scientific Hypotheses and Contributions

Large-scale medical imaging datasets are difficult to annotate manually due to the substantial time, cost, and clinical expertise required. To overcome this bottleneck, this research explores whether unsupervised annotation techniques can be used to organise a multimodal imaging dataset into semantically coherent groups. By extracting and combining features from raw DICOM images, structured metadata, and narrative diagnostic reports, the goal is to produce clusters that retain clinical meaning without the need for manual labelling, thereby enabling scalable dataset preparation for model pretraining.

Furthermore, this work examines whether CNN architectures pretrained on the RadiologyNET dataset offer advantages over models pretrained on general-purpose datasets such as ImageNet, or models trained from randomly initialised weights (*Baseline* models). The evaluation includes model generalisation, convergence speed, and stability, particularly in settings where the amount of annotated training data is limited or training resources are limited.

The key research hypotheses addressed in this thesis are as follows:

1. **Unsupervised annotation of medical data is a viable method of labelling medical data and grouping semantically similar data points together, facilitating the dataset's use for building foundation models.**

2. **Transfer learning from RadiologyNET foundation models can improve model performance, especially in medical tasks that are resource-scarce.**

In order to test these hypotheses, the following research aims are defined. First, the RadiologyNET dataset should be analysed, and semantically similar data points should be grouped together. Second, the generated pseudo-labels should be used for building RadiologyNET foundation models by pretraining several popular deep learning topologies as supervised pretraining tasks. Third, the impact of RadiologyNET pretrained weights should be evaluated on a variety of downstream tasks, with an evaluation extending to resource-limited tasks.

From the presented hypotheses and research aims, this thesis makes the following contributions to the field of transfer learning in medical machine learning:

1. A method for the automated grouping and labelling of semantically similar medical radiology images;

2. Development of RadiologyNET foundation models for transfer learning;

3. Evaluation of RadiologyNET foundation models against those pretrained on ImageNet and baseline models.

## 1.3. Research Methodology

The general research pipeline consists of four phases described below, with their general overview presented in Figure 1.1.

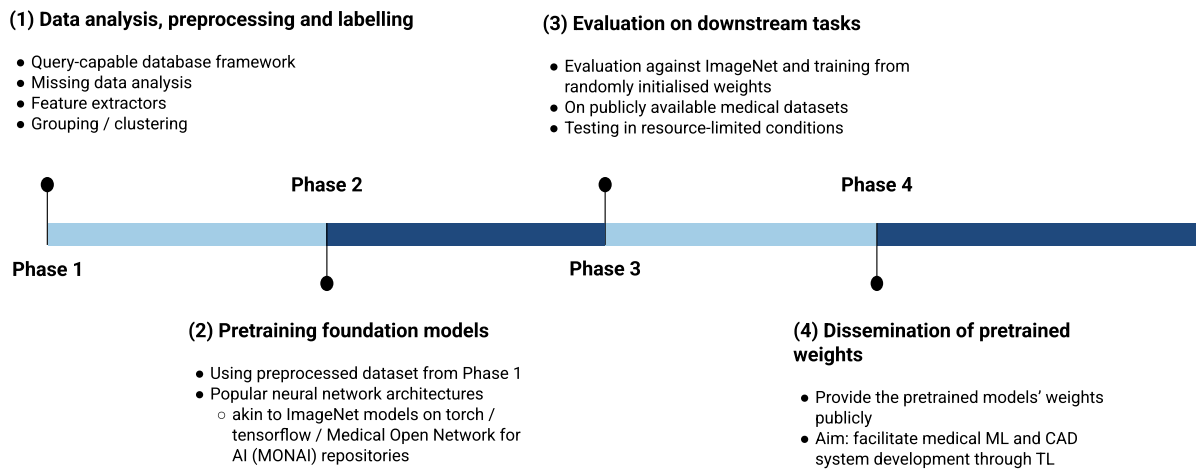



**Figure 1.1:** The general research phases and pipeline.

**The first research phase** involved a detailed analysis of the RadiologyNET dataset. Given the dataset's substantial size of approximately 13 terabytes, a query-capable database framework was established to enable efficient searching and analysis. This framework allowed for detailed examination of DICOM attributes, narrative diagnoses, and image content. From the full collection of approximately 25 million DICOM files, a high-quality subset was extracted based on defined criteria, filtering out incomplete, corrupted, or noisy data. Each data source was examined separately and subjected to dedicated preprocessing pipelines, including a missing-data analysis [49]. Appropriate feature extraction methods were applied to each data type to identify relevant patterns. These features were subsequently combined into unified embeddings, which were then grouped using unsupervised clustering techniques to automate the annotation process by producing pseudo-labels. This enabled the grouping of semantically similar examinations while segregating disparate ones, forming the basis for subsequent model pretraining.

**The second research phase** focused on building foundation models for TL. Several

widely used deep learning architectures were selected and pretrained on the RadiologyNET dataset, following a strategy analogous to the use of ImageNet [13, 14] in natural image processing. The pretrained models included InceptionV3 [50], VGG [51], DenseNet [31], various ResNet [30] configurations, EfficientNet [32], and MobileNet [52] variants. The objective of this phase was to produce models capable of extracting domain-relevant features from medical images.

**The third research phase** centred on evaluating the pretrained RadiologyNET models across a wide range of publicly available medical imaging datasets. The evaluation datasets were chosen to cover diverse imaging modalities (CT, MR, X-ray) and anatomical regions (extremities, lung, brain). The models were assessed on different tasks and task types, including:

- *Binary classification*, on the GRAZPEDWRI-DX [53] (detecting the presence of osteopenia in wrist radiographs) and COVID-19 [4, 5] (detecting COVID-19 in chest radiographs) datasets, with both datasets containing X-ray images;
- *Multiclass classification*, on the Brain Tumor MRI [54] dataset, where the goal is to classify between different types of tumours;
- *Regression*, on the Pediatric Bone Age Challenge [55] where the goal is to predict skeletal age from hand X-ray images;
- *Semantic segmentation*, on the LUng Nodule Analysis [56, 57] dataset, with the goal of segmenting lung nodules in CT images.

The performance of the RadiologyNET-pretrained models was compared against ImageNet-pretrained models and baseline models trained from randomly initialised weights. Each TL approach was tested across a range of hyperparameter values (e.g. learning rates), with multiple independent runs per each setting. Statistical tests were conducted to determine whether differences in performance across the three pretraining approaches were statistically significant.

**The fourth research phase** was dedicated to the dissemination of the RadiologyNET pretrained weights. These pretrained models were made publicly available to researchers and practitioners in the medical machine learning community at `https://`

`github.com/AIlab-RITEH/RadiologyNET-TL-models`. By providing access to domain-specific pretrained models, this work aims to contribute towards the development of medical machine learning systems and the overall understanding of TL in the medical domain.

## 1.4. Thesis Structure Overview

Chapter *1. Introduction* provides an introduction to the field of medical foundation models. The problem of annotated data scarcity in medicine is discussed, as are the approaches often used to mitigate this problem.

To make this thesis and the research presented here easier to follow, the subsequent chapters mirror the research phases shown in Figure 1.1. Specifically, chapter *2.* describes the first phase, and chapter *3.* the second phase. The third and fourth phases are addressed in parallel across chapters *4.* and *5.*.

– Chapter *2. Building the RadiologyNET dataset* refers to the first phase, introducing the RadiologyNET dataset in detail. From the entire dataset (which is substantial in size), a high-quality subset was extracted and pseudo-labelled using a fully automated process which requires no manual annotation. This chapter includes an extensive ablation study comparing and benchmarking different feature extractors for three data types (DICOM metadata, images and narrative diagnoses), detailing how each data source impacts the final pseudo-labels. The pseudo-labelled dataset is presented and discussed.

– Chapter *3. RadiologyNET Pretraining and Experimental Design* details the process of pretraining on the pseudo-labelled data. Multiple popular neural network architectures are pretrained to be used in medical downstream tasks. The experimental setup is described here, with a given overview of metrics and statistical tests to be used in analysing the performance on downstream tasks.

– Chapter *4. The Efficacy of RadiologyNET Foundation Models* describes the chosen medical downstream tasks as well as the rationale for choosing each task. The results of TL from RadiologyNET are presented and discussed in this chapter, as well as the performance of TL in resource-limited conditions.

– Chapter *5. Domain Influence on Performance and Interpretability* focuses on the impact of patterns learned during pretraining, e.g. whether pretrained weights have an influence on downstream model interpretability. Also, this chapter presents the influence that pretraining on specific medical imaging modalities might have on the performance on downstream tasks.

Chapter *6. Conclusion* summarises the research presented in this thesis, reiterating the limitations of this work and possible future research directions.

# 2. Chapter

# BUILDING THE RADIOLOGYNET DATASET

## 2.1. Dataset origin and structure

The dataset consists of approximately 2.3 million unique exams completed between 2008 and 2017 and performed through standard clinical practice at Clinical Hospital Centre Rijeka. The data were retrospectively collected in 2017 and anonymised during the extraction process to ensure the removal of all personally identifiable (sensitive) information. Ethical approval for data collection and processing was obtained from the relevant Ethics Committee. As mandated by the approval, the dataset must remain private and may not be publicly released in its current form.

Each examination could result in a respective diagnosis and at least one (often more than one) DICOM file, which was then stored at Clinical Hospital Centre Rijeka's Picture Archiving and Communication System (PACS). The diagnoses were written in the Croatian language, while each DICOM file was composed of two parts:

1. **the file body**, which contains pixel data (the actual radiology image),
2. **the file header**, which stores structured metadata describing the imaging context (e.g. modality of image acquisition, body part examined, imaging protocol). Example DICOM tags are given in Table 2.1.

An example examination is illustrated in Figure 2.1.

**Table 2.1:** Overview of selected DICOM tags, their data types, and common usage in medical imaging workflows.

| DICOM Tag | Data Type | Usage / Application |
|---|---|---|
| `BodyPartExamined` | Code String | Indicates the anatomical region imaged. Example values: `EXTREMITY`, `ANKLE`, `ABDOMEN` |
| `Modality` | Code String | Specifies the modality of image acquisition. Example values: `CT`, `MR`, `CR`, `RF`, `XA` |
| `StudyDescription` | Long string | Provides a brief textual summary of the imaging study. Example value: `T Ankle joint 2 views` |
| `ProtocolName` | Long string | User-defined description of the conditions under which imaging was performed. Example value: `T107aL Ankle joint a.p.` |
| `RequestedProcedure Description` | Long string | Institution-generated administrative description of the requested procedure. Example value: `RTG GLEZANJ` (en. *ankle radiograph*) |
| `WindowCenter` | Numeric (Float) | Represents the centre of the display intensity range; used in preprocessing and display windowing for consistent visualisation; e.g. can be used in CT images to highlight different anatomical areas (lung, soft tissue, bone...) |
| `WindowWidth` | Numeric (Float) | Defines the width of the display intensity range; used alongside `WindowCenter`. |
| `RescaleType` | Long string | Describes the units of pixels after applying the rescale step. Example value: `HU` (Hounsfield Units, used in CT images) |
| `RescaleIntercept` | Numeric (Float) | Used to rescale the image into units specified by `RescaleType` |
| `RescaleSlope` | Numeric (Float) | Used to rescale the image into units specified by `RescaleType`; used alongside `RescaleIntercept`. |

## 2.2. Query-capable database framework

The total number of DICOM files obtained from Clinical Hospital Centre Rijeka PACS reached approximately 25 million [41] (24,969,869, to be exact), with $\approx$ 13 terabytes of space required to store the entire dataset (images and diagnoses). Working with large volumes of DICOM data introduces inherent complexities, primarily stemming from the amount of time required to read and process all files. In addition to the number of files involved, the sheer size of DICOM files also poses a challenge, as the size can range from several kilobytes to hundreds of megabytes in size, especially in the case of high-resolution

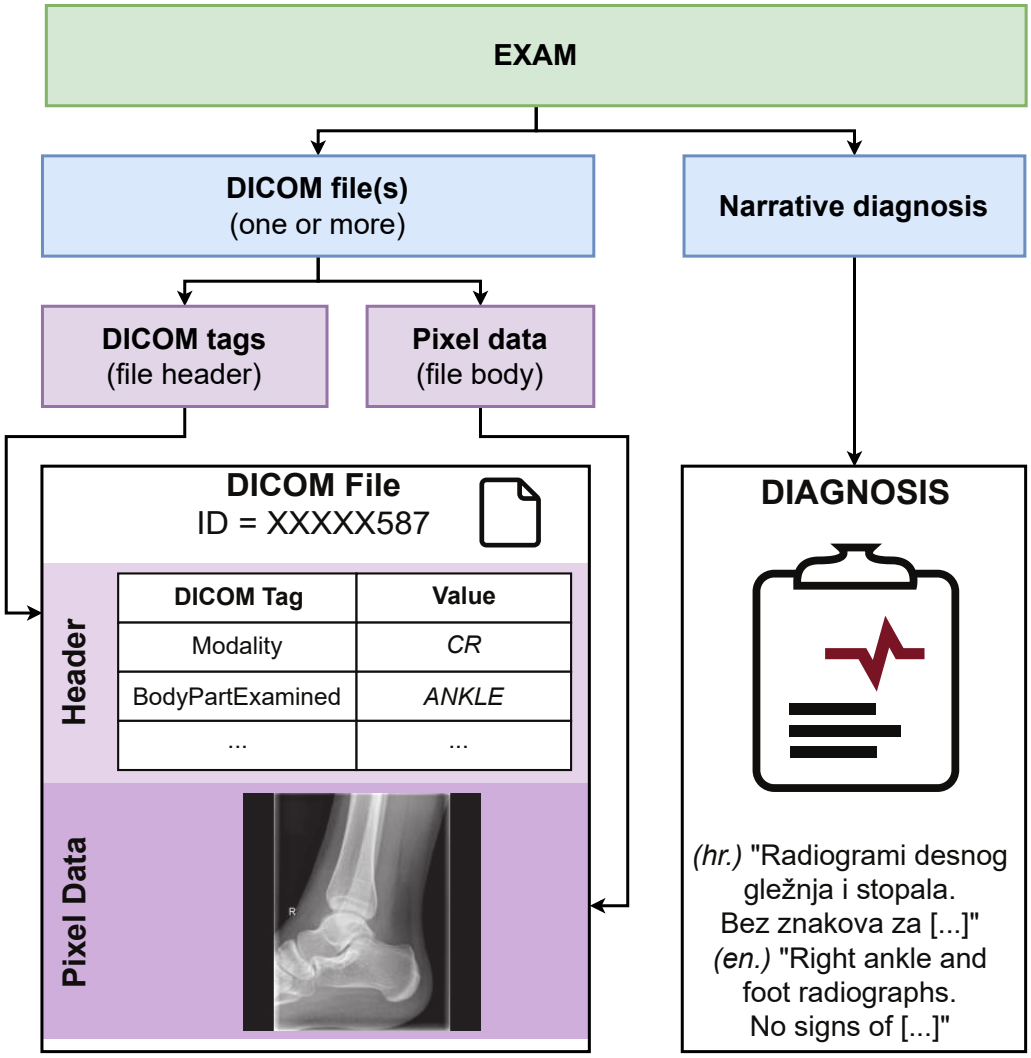



**Figure 2.1:** Structural overview of a single examination in the RadiologyNET dataset. The example diagnosis was originally written in Croatian; an English translation is included in the figure for illustrative purposes.

images or multi-frame studies. For this reason, a query-capable framework was built, to allow easy, efficient and fast retrieval (and filtering) of data points.

The folder structure of raw, gathered images is as follows. Each DICOM file is named after its unique identifier from Rijeka PACS, for example, the DICOM file whose ID is *12345678* is named `12345678.dcm`. The root directory `issa/` contains 28,660 subfolders, where each subfolder contains up to a thousand DICOM files. Each subfolder is named as the prefix of the DICOM files stored inside it. For example, if a DICOM file is named `12345678.dcm`, then it will be saved at `issa/12345/12345678.dcm`. The directory `12345/` is named as the DICOM filename prefix, containing all but the last three digits. An

illustration of the full directory structure is given below.

```
issa/
├── 10000/
│   ├── 10000000.dcm
│   ├── 10000001.dcm
│   ├── ...
│   └── 10000999.dcm
├── ...
├── 12345/
│   ├── 12345000.dcm
│   ├── ...
│   └── 12345999.dcm
├── ...
└── 9999/
    ├── 9999000.dcm
    ├── 9999001.dcm
    ├── ...
    └── 9999999.dcm
```

The two-phase process of building the query-capable database framework is shown in Figures 2.2 and 2.3. In the first phase (shown in Figure 2.2) the contents of each `issa/` subfolder were parsed using the `pydicom` [58] library, which is a python package used in DICOM file reading and processing. For each *readable* DICOM file, metadata attributes were extracted and stored in a Comma-Separated Value (CSV) file. During this process, corrupted files or those with entirely empty metadata were excluded from further processing. Additionally, attributes capturing the structure of the pixel data, such as *PixelDataShape* and a boolean *IsRGB* indicator, were computed based on the shape of the pixel data found in the file body. These were stored alongside standard DICOM tags to enable filtering of red-green-blue (RGB) images, and multi-frame/volumetric studies, which typically differ in dimensionality from (for example) single-frame CR images. Each subfolder yielded one CSV file containing one row per DICOM file. The output CSV adopted the name of its corresponding folder. For example, a folder named `12345/` containing three DICOM files would result in a file `12345.csv` with three metadata rows. In total, 25,632 such CSV files were produced, a number lower than the original 28,660 folders due to the exclusion of empty folders or folders containing only invalid files.

In the second phase (shown in Figure 2.3), the exported CSV files were subsequently aggregated into a single metadata table using the `dask` python package [59], which is

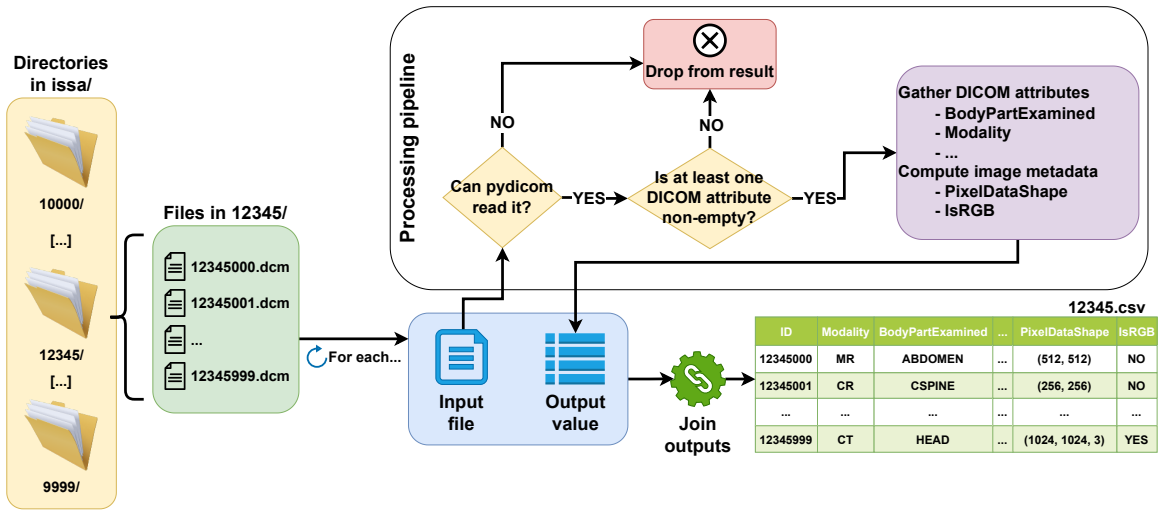



**Figure 2.2:** The process of reading DICOM files and exporting metadata into CSV files.

typically used for parallelised reading and processing of tabular data. The combined table was then analysed to remove non-informative columns. The DICOM metadata content differs depending on the imaging modality used to capture the image – for example, *PlateID* is a DICOM attribute which is present only in CR images. With this in mind, sample-wise selection of useful DICOM attributes was impossible, requiring the selection process to be performed on the entire dataset. Thus, all of the table's columns were analysed to see if they contain useful information (or rather, if they contain any at all). Any column containing empty data throughout all of the DICOM instances was dropped from the table completely. However, if a column had at least one non-empty value in any row, then it remained in the table.

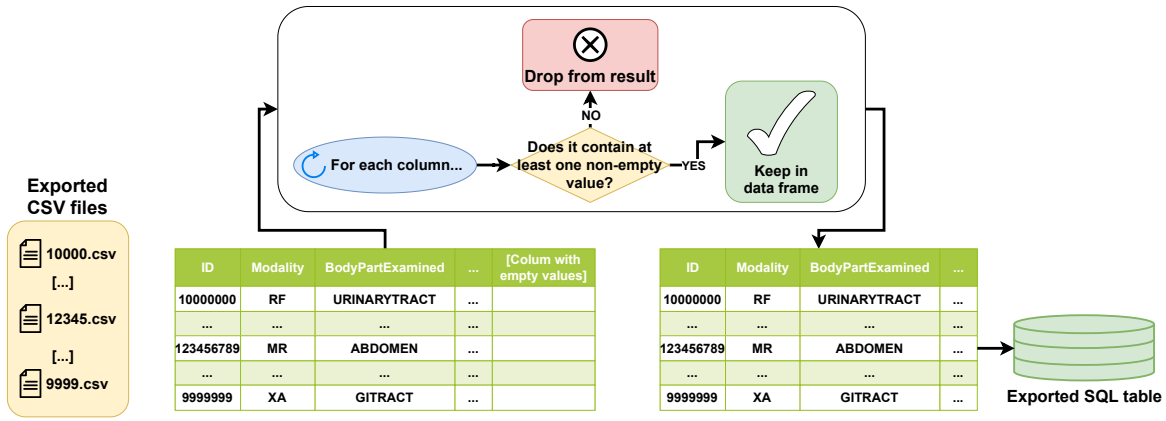



**Figure 2.3:** The process of joining multiple CSV files, dropping empty columns, and exporting into the final result.

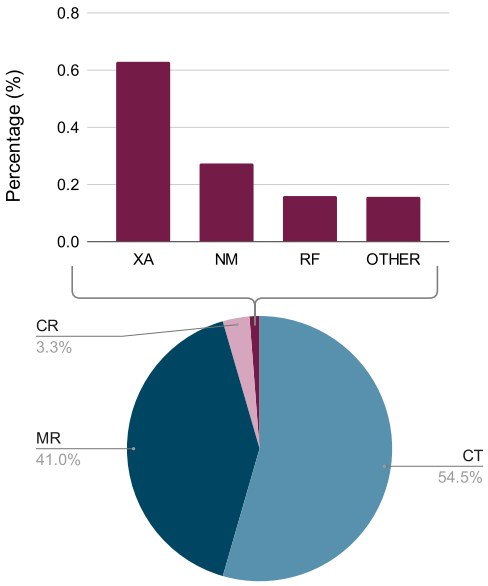



**Figure 2.4:** Distribution of imaging modalities in the RadiologyNET dataset [41].

The table was stored as a Structured Query Language database table. The exported file requires approximately 85 gigabytes of space and contains queryable information of 23,753,712 distinct DICOM files (i.e. rows) with 653 DICOM tags (i.e. columns), each with at least one non-empty value. Once the query-capable metadata framework was established, it served as the foundation for subsequent analysis, data sampling and feature extraction.

## 2.3. Data analysis

The most prevalent modalities in the dataset were CT, MR, XA, NM (Nuclear Medicine), and RF, as shown in Figure 2.4, with 95.5% of the entire dataset being comprised solely of CT and MR images. A majority of the images, comprising 99.35% of the dataset, were in greyscale format, while only a small portion (0.65% of the data) consisted of RGB images, captured mostly using CT (70.9% of cases) and MR (27.9% of cases) imaging techniques. Although radiology images are mostly captured and stored in grayscale, RGB images can be a result of overlaying colour-mapped positron emission tomography on top of other image sources to allow a clear correlation between anatomical and functional imaging [60]. Examples of images captured across different modalities (CR, CT, MR, XA, and

RF) are shown in Figure 2.5. There were 75 distinct values in the *BodyPartExamined* tag. Among these categories, the most commonly examined body parts in the dataset were the Abdomen (24.5%), Head (23.2%), Chest (16.5%), and Breasts (6.9%). The *BodyPartExamined* tag was left empty in approximately 8.4% of the instances, suggesting either missing or unspecified information regarding the examined body part.

Not all of the 2.3 million examinations contained an associated diagnostic report; approximately 6.96% of entries had an empty or missing diagnosis, while 93.03% included a non-null, non-empty value. Upon further inspection, it was discovered that some of the non-empty diagnoses were less than 5 characters long. These were presumed to be anomalies as such short diagnoses could seldom carry useful information. Empty diagnoses and diagnoses which had less than 5 characters were excluded from the further use.

Some of the examinations contained over 1,000 associated DICOM files. This can occur when radiologists use multiple projections or views, or in continuous image capture – such as in flouroscopy. As flouroscopy captures real-time moving images (i.e. time series), the Clinical Hospital Centre Rijeka PACS system would sometimes store them as separate DICOM files instead of storing them in a single multi-frame file. This behaviour varies depending on the imaging device and its DICOM implementation.

### 2.3.1. Data subsampling

Each data point was defined as a unique triplet comprising a DICOM image, its associated DICOM metadata (tags), and the corresponding narrative diagnosis. An example of two data points extracted from a single exam is illustrated in Figure 2.6. In this figure, two data points were extracted during the same examination, therefore the images were were associated with the same diagnostic report. The diagnosis was originally written in Croatian (*hr.*), and a portion of this diagnosis is displayed in the figure, along with its English (*en.*) translation.

From the full dataset, a subset of 135,775 DICOM files and corresponding narrative diagnoses was sampled for building the unsupervised annotation pipeline. As there were over examinations with 1,000 associated DICOM files, special attention was given to include as many complete diagnoses as possible during the sampling process. To maintain balance between the number of images and distinct diagnoses, a maximum of 15 DICOM

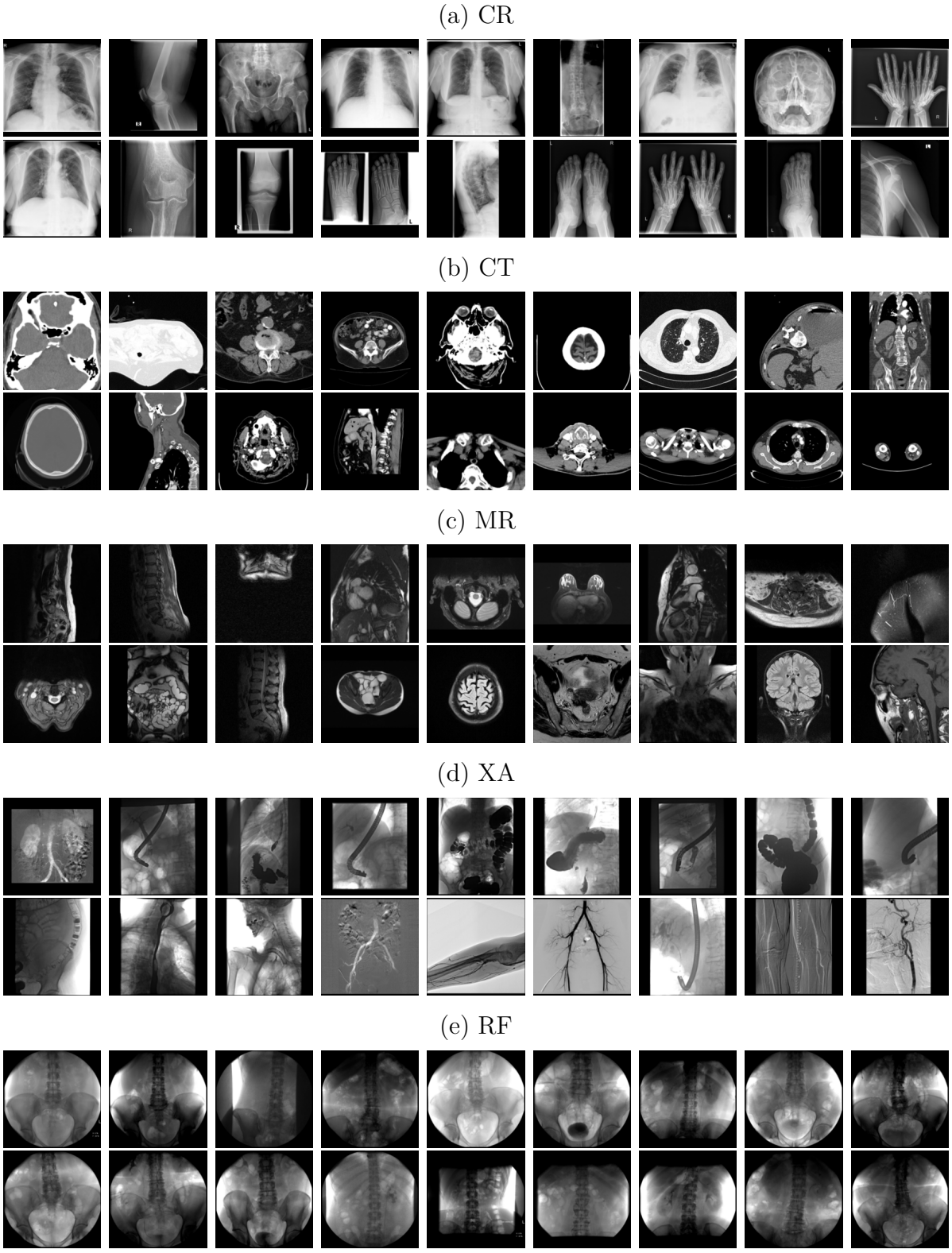



**Figure 2.5:** Examples of images found in the RadiologyNET dataset. Each subfigure represents images captured in the specified modality.

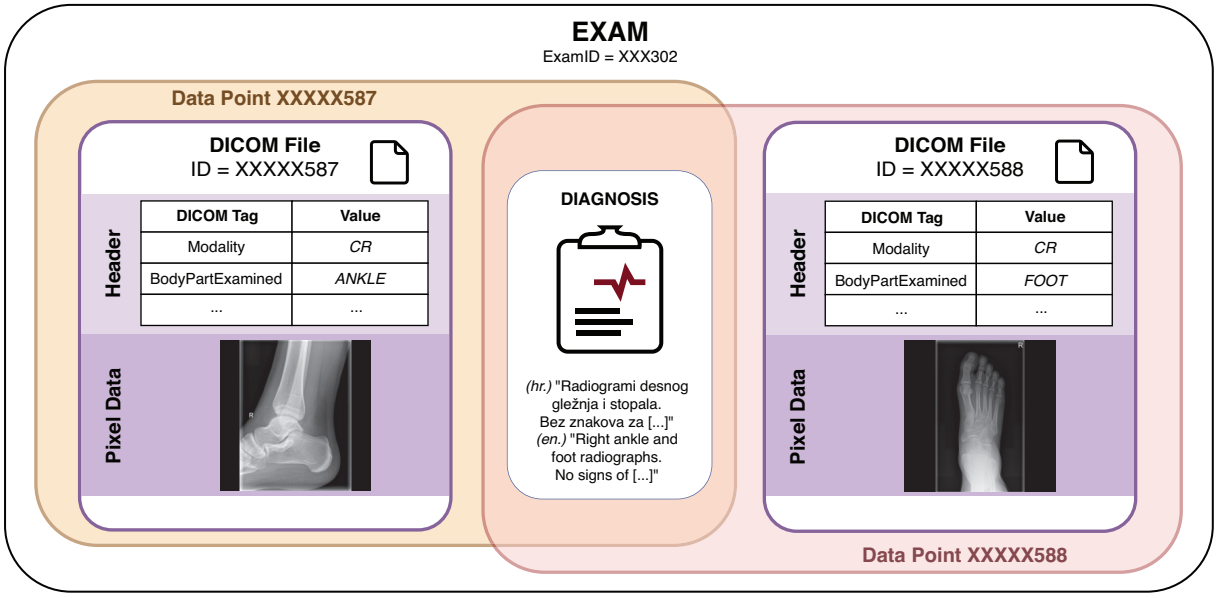



**Figure 2.6:** An example of two data points obtained from a single examination, which included two CR images of the right ankle and foot [41].

**Table 2.2:** The sizes of train, test and validation subsets used for building the unsupervised annotation pipeline [41]

| Subset | Exam (diagnoses) count | DICOM file count |
|---|---|---|
| Train | 50,528 (80.00%) | 108,542 (79.94%) |
| Validation | 6,316 (10.00%) | 13,596 (10.01%) |
| Test | 6,316 (10.00%) | 13,637 (10.05%) |
| **Total** | 63,160 (100.00%) | 135,775 (100.00%) |

files per examination was allowed; exams exceeding this threshold were excluded, i.e. there were no examination where more than 15 data points could be extracted. The resulting subset included images from five modalities: CT, MR, CR, XA, and RF. Although initially included in the subsampling process, images from the NM modality were later removed due to frequent association with missing or low-quality (short, uninformative) diagnostic reports. Table 2.2 shows the extract sizes of each sampled subset.

### 2.3.2. Missing data and preprocessing

A general pipeline of each data source (DICOM tags, images, diagnoses) preprocessing is shown in Figure 2.7. In short, each of the three data sources required a distinct preprocessing approach, for example, DICOM tags required additional filtering and missing data analysis. Images come in different radiology imaging modalities, each requiring specific preprocessing steps; and are often stored in different bit-depths, thus needing ad-

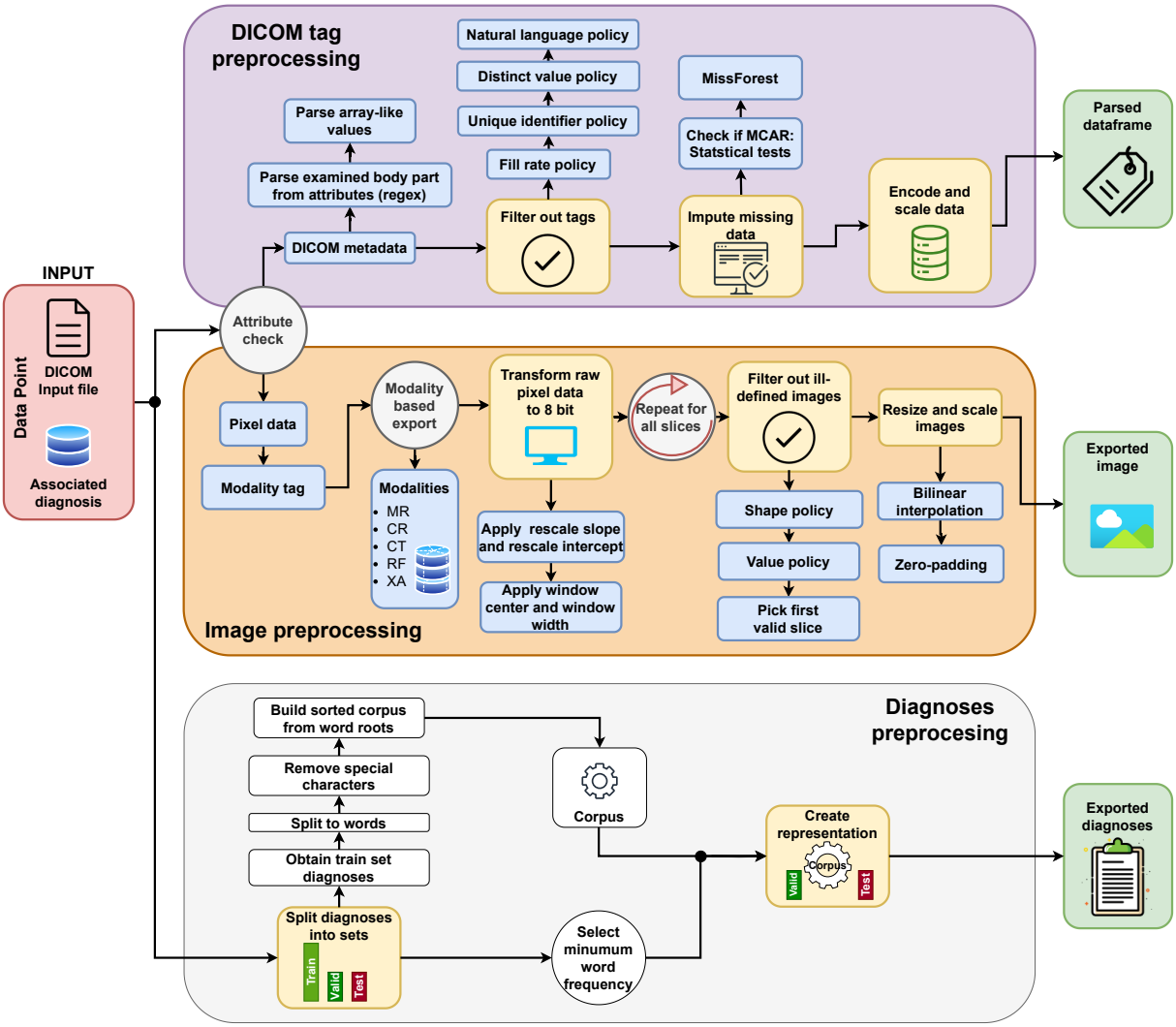



**Figure 2.7:** The process used for exporting images, DICOM tags and narrative diagnoses [41].

ditional scaling and resizing. On the other hand, medical diagnoses were written as free text, therefore the words were stripped to their roots to capture the core meaning. To create a representation of the diagnosis, text-processing algorithms often require a corpus based on frequently used words, which was built using the training set (Table 2.2). The preprocessing of each data source is described in more detail in the following subsections.

### DICOM tags

The first encountered problem was **DICOM tags with missing *BodyPartExamined***, which contained an empty value in 59.4% cases. On the other hand, tags such as *ProtocolName*, *StudyDescription* and *RequestedProcedureDescription* (described in Table 2.1) faired better, having empty values in only 10.9%, 7.39% and 33.8% of instances, re-

spectively. Wherever *BodyPartExamined* was empty, at least one of the mentioned tags contained a value from which one can infer the examined body part, which is why these three particular tags were chosen. In order to solve the missing values for *BodyPartExamined* tag, there were 53 regular expressions written, a snippet of which is shown in Code Listing 2.1. These regular expressions contained rules for imputing *BodyPartExamined* from the *ProtocolName*, *StudyDescription* and *RequestedProcedureDescription* tags, and were written in a way that accounts for possible typographical errors (e.g. *torax* and *thorax*), multiple languages used by physicians (e.g. Latin: *calcaneus*; English: *heel bone*; and Croatian: *petna kost*), possible abbreviations (e.g. cervical spine: *c-spine*, *c_spine*, or *cspine*), and which procedures impact which body part (e.g. *chemoembolization* is tied to the liver). These rules were written under a radiologist's guidance, as there is no straightforward ruleset for perfect *BodyPartExamined* mapping.

**Code Listing 2.1:** Regex-based mappings for inferring `BodyPartExamined` (BPE)

```
BPERegex(to_val='WRIST', regexpr='wrist|rucni')
BPERegex(to_val='FINGER', regexpr='finger|prst')
BPERegex(to_val='ARM', regexpr='forearm|upper extrem|gornj(?:i|ih|eg)
        ekstrem|(?:nad|pod)laktic|humerus|ruk(?:a|e|u)')
BPERegex(to_val='HAND', regexpr='saka|sake|hand')
BPERegex(to_val='ELBOW', regexpr='lakat|elbow')
BPERegex(to_val='HEEL', regexpr='calcaneus|heel|petn(?:a|e)|peta')
BPERegex(to_val='ANKLE', regexpr='glezanj|gleznja|ankle|skocni')
BPERegex(to_val='KNEE', regexpr='knee|koljen')
BPERegex(to_val='FOOT', regexpr='stopal|foot')
BPERegex(to_val='LEG',
    regexpr='lower (?:leg|limb|extrem)|nog(?:a|e|u)|femoral(?:|ne) art|do
        (?:|lj)nj(?:i|ih|eg) ekstrem')
BPERegex(to_val='CSPINE',
    regexpr='vratna(?: |_|-)kra|c(?:-|_|)spine|cervikalne kralj')
BPERegex(to_val='LSPINE', regexpr='lumbaln(?:a|e)|l-spine|l_kralj')
BPERegex(to_val='TSPINE', regexpr='torakaln(?:a|e) kralj|t-spine')
BPERegex(to_val='CHEST',
    regexpr='thorax|lung|torax|toraks|src(?:a|e)|sternum|ribs|rebra|pluc(?:
        a|na)|myocardial|cardiac|subklavij|pulmonary|pulmonal|lung|heart')
BPERegex(to_val='HIP',
```

```
    regexpr='sa kukom|hip joint|(?:desni|lijevi|rtg|rdg) kuk')
BPERegex(to_val='PELVIS',
    regexpr='pelvis|zdjelic|ilijak|iliac|sacrum|si zglob')
BPERegex(to_val='ABDOMEN', regexpr='abdom(?:en|inal)')
BPERegex(to_val='WHOLEBODY',
    regexpr='cijelo tijelo|cijelog tijela|whole body|wholebody|total body')
```

The final result of *BodyPartExamined* imputation (based on knowledge, decision rules, and regular expressions) resulted in an increase from 40.6% to 100% non-empty instances. However, it should be noted that there was still a possibility of erroneously imputing *BodyPartExamined* from other tags. Specifically, DICOM tags *StudyDescription*, *ProtocolName* and *RequestedProcedureDescription* are input manually by a performing physician. As such, other than typographical errors, it is possible that other types of errors could lead to mislabelling of a body part that was not accounted for. However, these cases were presumed to be anomalies and only present in a few DICOM files.

The next group of tags needing additional care are stringified arrays - **DICOM tags with multiple values**. Some DICOM tags can contain multiple values, for example, *WindowWidth* and *WindowCenter* can have a single value such as `"1134"`, but also multiple values, e.g. `"[1134, 423]"`, which can occur when the performing physicians wants to highlight multiple tissue types within the same image. Such tags were parsed from a single stringified array-like tag into multiple tags, which resulted in *WindowCenter* dissolving into *WindowCenter0* and *WindowCenter1*, etc.

Another challenge was **selection of appropriate DICOM tags**. There were 654 different DICOM tags that appeared at least once in the whole subset. However, many proved to be uninformative due to either being empty in most instances or having only one distinct value. A fill rate threshold was imposed on each tag, and each tag with less than 35% non-empty values was removed. Furthermore, all DICOM tags with less than 2 distinct values were discarded, along with tags containing unique identifiers, such as *SOPInstanceUID*. After this, continuous and categorical DICOM tags were separated, and categorical variables were further examined. In particular, some of the tags contained natural language, which fell out of the scope of DICOM tag processing. The eliminated tags include the aforementioned *ProtocolName*, *StudyDescription* and *RequestedProce-*

*dureDescription*, accompanied by *AdmittingDiagnosesDescription*, *ImageComments*, etc. The remaining categorical variables had no more than 50 unique values. After this, 55 tags remained, of which 28 were continuous and 27 were categorical variables.

The final problem to solve regarding DICOM tags was **missing data analysis**. As was mentioned before, *BodyPartExamined* can be directly imputed from other tags via regular expressions, but other values' imputation is not as straightforward. Before imputing these values, the DICOM tags with missing data were analysed further, to test whether their *missingness* was completely random, or there is a pattern from which it is possible to infer the missing values [61]:

- *Missing-at-completely-random*: the probability of a value being missing is independent of both observed and unobserved data. In this case, the distributions of observed and missing values are expected to be similar.
- *Missing-at-random*: the probability of a value being missing depends only on observed data. Differences between missing and observed values can be accounted for using other variables.
- *Missing-not-at-random*: the probability of a value being missing depends on unobserved data (including the missing value itself), and thus the difference between missing and observed values cannot be accounted for by other variables.

*Missing-at-completely-random* and *missing-at-random* data can be imputed without introducing bias. However, data that are *missing-not-at-random* are more challenging, as the mechanism behind the missingness cannot be addressed using observed variables, making unbiased imputation generally impossible. While *missing-at-random* and *missing-at-completely-random* can be statistically tested, there is no statistical test that can conclusively determine whether a variable is *missing-not-at-random*, as this would require access to unobserved values.

To determine if data were *missing-at-completely-random* or *missing-at-random* [61, 62], univariate statistical tests were performed as described by Enders [49]. Statistical tests differed based on whether the observed variable was discrete or continuous and, in the latter case, if it was normally distributed. If a continuous variable was normally distributed (Shapiro-Wilk, $p \geq 0.05$), then an Independent samples t-test was performed,

while a Mann-Whitney U was applied otherwise. In the case of categorical data, a $\chi^2$ (chi-square) test was used. A variable would not be considered *missing-at-completely-random* if its missingness influenced the distribution of at least one other variable, i.e. there was a statistical difference in the distribution where said variable was missing versus where it was not. Although the used approach has its drawbacks if multivariate interactions exist (and Little's *missing-at-completely-random* test can be more appropriate) [49], it can bring attention to dependencies between variables. Consulting with the DICOM standard [42, 63] strengthens the assumption that data is likely *missing-at-random* and not *missing-at-completely-random*, as most of the DICOM tags depend on the modality used to capture the image. The missing values were imputed using MissForest [62, 64], which uses random forests to impute missing data and had been previously shown to work well with *missing-at-random* data [65]. Mean Squared Error (MSE) was used as the criterion for continuous and Gini impurity for categorical variables. MSE is calculated as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2.1}$$

where $n$ is the number of instances, $y_i$ is the ground truth value for the $i$-th sample, and $\hat{y}_i$ is the predicted value. Gini impurity is calculated as:

$$\text{Gini}(t) = 1 - \sum_{i=1}^{C} p_i^2 \tag{2.2}$$

where $C$ is the number of classes and $p_i$ is the proportion of instances of class $i$ in node $t$. Node $t$ is a node in a decision tree from the MissForest algorithm; a lower Gini impurity indicates a *purer* node, which in turn leads to more reliable imputations for the missing categorical values. After imputation, the categorical variables were one-hot encoded, and continuous variables were scaled to fit the range $[0.00, 1.00]$.

### Images

In order to export images from DICOM files, the presence (and the corresponding values) of several DICOM tags must first be checked, as certain modalities require different preprocessing techniques. Medical radiology imaging does not often capture ready-to-view images and is highly dependent on the diagnostic context. An example of context-

depending imaging is CT, where images differ significantly based on the observed tissue (e.g. bone versus soft tissue). This means that the stored pixel values have to be transformed into display-ready values, and there are several steps to achieving this: (i) rescaling the image, and (ii) windowing the image. For example, the pixel intensity of images can be scaled to map raw image values to ranges that preserve (or enhance) diagnostically relevant regions [66, 67]. As a first step in this process, the values of DICOM tags *RescaleSlope* and *RescaleIntercept* are checked and used to scale the raw pixel image $x_{Ir}$ using the formulation:

$$x'_I = R_s \cdot x_{Ir} + R_i \tag{2.3}$$

where $x_{Ir}$ and $x'_I$ are raw pixel values (the input image) and the rescaled pixel values, respectively; $R_s$ is the rescale slope, and $R_i$ is the rescale intercept. If *RescaleSlope* and *RescaleIntercept* are not available in DICOM metadata, then the default values $R_s = 1$ and $R_i = 0$ are used, which essentially preserve the original pixel values (i.e. these parameters leave the pixels unchanged).

The equation 2.3 assumes that the transformation is linear. However, there are cases where DICOM metadata contains look-up tables for non-linear transformations, for example, there is a DICOM tag called *Modality LUT Sequence*. In cases where the transformation from stored pixel values to meaningful output requires the use of a look-up-table (e.g. *Modality LUT Sequence*) – these data points were excluded from further processing. Implementing LUT-based transformations is context-dependent and may vary across modalities (and capturing devices / vendors) [68].

DICOM images are captured in various bit depths. For example, CR images can be captured in 10-, 12- or 16-bit depth [67], and different bit depths lead to a different range of possible pixel values. A 10-bit image can have pixel values in the range [0, 1023], while a 16-bit image can contain pixel values in the range [0, 65,535]. As machine learning algorithms and neural networks are sensitive to data ranges, the images were mapped to 8-bit depth to mirror ImageNet's bit depth (i.e. range $[0, 255]$) and stored in Portable Network Graphics (PNG) format. Although scaling with the maximum pixel value is a valid solution as it essentially normalises the image, this can lead to a loss of clinically relevant information (and preservation of irrelevant information). For this reason, there

is an established practice in medical radiology, where the performing physician can determine the clinically relevant pixel values and store them in DICOM as *WindowCenter* and *WindowWidth* tags. These *windowing* parameters can be used to remove the clinically irrelevant pixel intensities, thus reducing the necessary bit depth required to store the image. This process is shown in Figure 2.8, where the original image (on the left) is in 12-bit depth and the output image (on the right) is the windowed image in 8-bit depth. As standard common-purpose displays cannot show more than 8-bit colour depth, for visualisation purposes, the image on the left was scaled via division with the maximum possible pixel value. The differences are noticeable: the windowed image (on the right) highlighted the bones and removed irrelevant shades of gray.

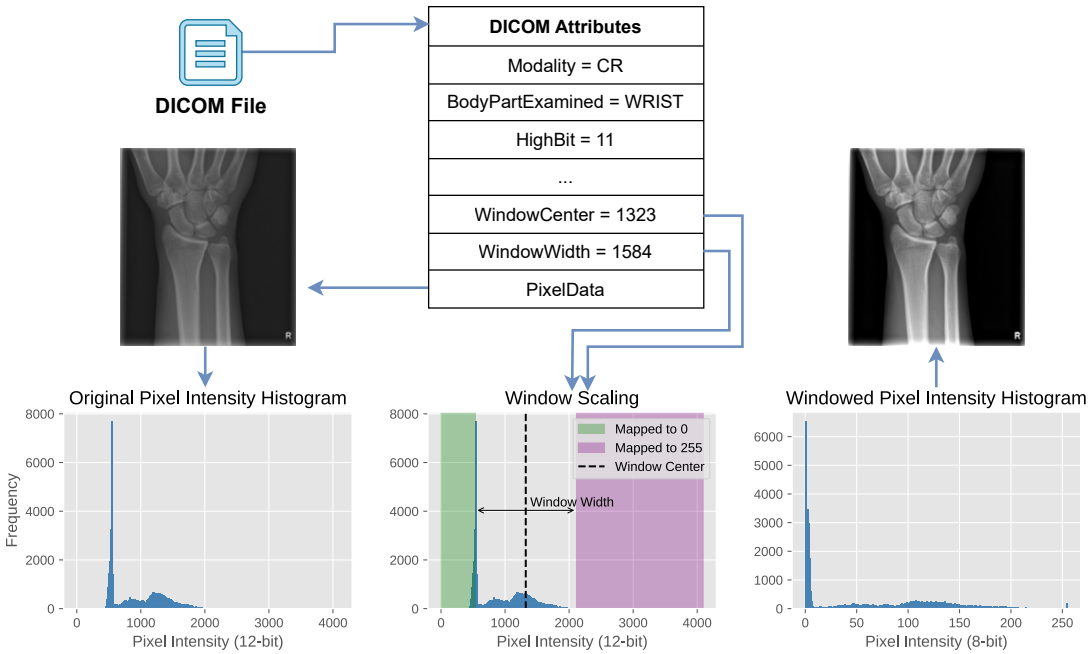



**Figure 2.8:** The process of *windowing*, which uses the DICOM tags *WindowCenter* and *WindowWidth* to remove uninformative pixel values from the pixel value histogram. This process follows the Equation 2.4 [67].

The windowing transformation is defined as follows [66, 67]:

$$x_I = \begin{cases} 0, & \text{if } x'_I \leq W\text{min} \\ 255, & \text{if } x'_I \geq W\text{max} \\ \left(\frac{x'_I - W\text{min}}{W\text{max} - W\text{min}}\right) \cdot 255, & \text{otherwise} \end{cases} \tag{2.4}$$

where:

- $x'_I$ is the rescaled pixel intensity after applying rescale slope and intercept as given in Equation 2.3;
- $x_I$ is the final 8-bit output pixel intensity;
- $W_{\min}$ and $W_{\max}$ represent the lower and upper bounds of the window, calculated as:

$$W_{\min} = W_c - \frac{W_w}{2}, \qquad W_{\max} = W_c + \frac{W_w}{2} \tag{2.5}$$

Here, $W_c$ is the window centre, and $W_w$ is the window width. The piecewise function ensures that pixel values below $W_{\min}$ are clipped to 0, and those above $W_{\max}$ are clipped to 255, with linear scaling in between.

*WindowCenter* ($W_c$) and *WindowWidth* ($W_w$) are read out from the DICOM metadata. Some modalities can have multiple windowing values present, e.g. a physician might want to highlight parts of the bone tissue and parts of the soft tissue in a single CT image. This can result in multiple pairs of *WindowCenter* and *WindowWidth* values in the DICOM header. In such cases, only the first *valid* value was used. The windowing parameters were considered valid if the resulting image was not entirely single-coloured (e.g. completely black or completely white). For multi-slice imaging modalities such as MR (where the image can essentially be a volume), only the first slice containing diagnostically meaningful data was used. Slice validity was determined using two filtering policies:

- *Value policy*: The ratio $r_V$ of distinct pixel values to the maximum possible number of distinct values was computed, and a slice was deemed *valid* if $r_V > t_V$. The threshold $t_V = 0.1$ was determined empirically through trial-and-error, by observing the pixel histograms of images dropped by the policy. The *value* policy resulted in the removal of images that contained very few distinct shades, i.e. images that were not completely single-coloured, but were *almost* single-coloured.
- *Shape policy*: Some DICOM files were possibly corrupt or erroneously saved, leading to images not being stored as *images*, but as *vectors*. To exclude non-image content (e.g. 1D vectors), the ratio $r_S$ of image width to height was computed, and a slice

was retained only if $r_S > t_S$. The threshold $t_S = 0.1$ was determined empirically through trial-and-error, by observing the shapes of images dropped by the policy.

These filtering criteria ensured that only informative images were included in the exported dataset.

In summary, the raw pixel data is subjected to a three-step transformation. First, intensity values are adjusted using the *RescaleSlope* and *RescaleIntercept* parameters to convert raw pixel values into meaningful intensity units. Second, the rescaled values are mapped to the standard 8-bit display range using the *WindowCenter* and *WindowWidth* parameters, and then stored in PNG format. Finally, the uninformative images were removed through applying the following criteria: the *value policy*, which ensures sufficient pixel intensity variation, and the *shape policy*, which filters out improperly shaped or non-image slices. The remaining images were resized to $256 \times 256$ pixels and zero-padded where necessary to preserve aspect ratio.

**Narrative diagnoses**

The preprocessing pipeline for narrative diagnoses is shown in Figure 2.7. These diagnoses, written in Croatian, describe patient conditions and medical findings. In comparison to the English language, the Croatian language differs as it has grammatical cases and verb suffixes, meaning that the same word can be written in different forms. By using different grammatical cases or changing the verb suffix, the entire context of a sentence can be changed; and vice-versa, words with the same semantic meaning can appear in different forms. To address this, the initial step involved reducing words to their root forms, to better capture their core meaning and to allow grouping based on semantic similarity.

Diagnosis preprocessing was carried out in several stages. First, (i) all diagnoses in the training set were tokenised into individual words, while removing special characters such as commas, colons, and semicolons. Next, (ii) words were stemmed using a rule-based Croatian stemmer developed by Ljubešić et al. [69]. Then, (iii) a vocabulary of 54,790 unique words was compiled from the training subset, which includes all words that appeared at least once. Upon a closer inspection of the obtained corpus, it was observed that there were words which appeared only once, and were often the result of

typographical errors made by physicians during manual entry. Hence, to improve the generalisation capabilities of text-based models, (iv) a parameter which regulates the minimal number of occurrences (i.e. minimum frequency threshold) was introduced to exclude such anomalies from the corpus.

## 2.4. Unsupervised annotation

Each data source presented limitations in terms of semantic clarity, completeness, or consistency. For example, inconsistencies were observed in metadata fields such as *BodyPartExamined*, where identical anatomical regions were labelled with varying specificity, such as *FOOT* and *LEG*, or more generic entries like *EXTREMITY* or *EXTREM*. Despite using regular expressions (an example of which were given in Code Listing 2.1), sometimes there were no ways to further define anatomical regions if the given value was non-specific. The example shown in Figure 2.6 shows an examination where both images contain valid *BodyPartExamined* values (each correctly referencing the observed body part); however, the accompanying narrative diagnosis refers only to the combined set of images (*"ankle and foot radiographs"*). Had the *BodyPartExamained* been set as *EXTREMITY* in both cases, it would be impossible to discern which image depicts which body part based solely on the narrative diagnoses, thus requiring manual inspection (which, in a dataset of this scale, would not be feasible). Although the example diagnosis at least mentions the general antomical area (*"ankle and foot radiographs"*), there were diagnoses lacking any anatomical or procedural reference altogether (e.g. *"Kontrola nakon mjesec dana, bez vidljivih promjena"*, en. *"Regular check-up after a month, no visible changes"*; or *"Bez znakova koštane destrukcije"*, en. *"No signs of bone destruction"*).

Given the absence of consistent and complete ground truth annotations, none of the available data sources (DICOM metadata, image data, or narrative diagnoses) were individually suitable for direct use as supervisory labels. Instead, the unsupervised annotation pipeline was designed to identify and exploit latent patterns in the data to group semantically similar examinations and (potentially) identify outliers and anomalies.

This section is structured as follows. First, the feature extraction strategies applied to each data source are described in detail. Each data source was processed independently using type-specific techniques to capture its underlying structure and semantic content.

The extracted features were then combined and used to group data points based on similarity, thereby generating pseudo-labels through unsupervised clustering. To assess the contribution of each data source to the quality of the resulting groups, an extensive ablation study was conducted, evaluating the effect of including or excluding individual data sources. Finally, the composition and characteristics of the generated pseudo-labels are presented.

### 2.4.1. Feature extraction methods

After preprocessing, each data source underwent feature extraction, with different feature extraction techniques being applied depending on the data type.

**The DICOM tag feature extraction process** was performed using two dimensionality reduction techniques: principal component analysis (PCA) [45] and autoencoders (AEs) [44, 46]. For each method, an extensive grid search was conducted over their respective hyperparameters. Multiple AE configurations were trained, varying in learning rate, encoder–decoder architecture, and bottleneck dimensionality. All AE encoders comprised three fully connected (dense) layers with progressively decreasing dimensionality, each followed by a rectified linear unit (ReLU) activation [70]. The encoder output was passed through a central bottleneck layer, which served as the low-dimensional latent representation. The decoder mirrored the encoder layout. Model training was performed using MSE (Equation 2.1) as the loss function, Adam [71] as the optimiser, a mini-batch size of 32, and a maximum of 100 epochs. An early stopping criterion was applied, halting the training process if validation loss failed to improve for 5 consecutive epochs. Hyperparameter value ranges used in the grid search are summarised in Table 2.4. Although a wide range of learning rates was initially explored, empirical results from the first few hundred trained models indicated that $10^{-2}$ and $10^{-3}$ yielded lower reconstruction error.

**To extract features from image data**, several deep learning architectures widely used in medical image analysis were evaluated [72]. The candidate models included a convolutional autoencoder (CAE), the original U-Net [73], the recurrent residual U-Net (R2U-Net)[74], and an attention-enhanced U-Net (AttU-Net)[75]. These architectures were selected for their effectiveness in learning spatially coherent representations and their proven performance across various medical imaging tasks [72]. All models were trained

using mini-batches of size 32, and validation was performed two times per epoch to monitor convergence given the large volume of image data. The Adam [71] optimiser was used alongside MSE as the loss function. Models were trained for a maximum of 40 epochs, with early stopping applied if the validation loss did not improve over 5 consecutive validation steps. To see whether any further dimensionality reduction could improve clustering results, PCA was applied with an extensive grid search of hyperparameters, as shown in Table 2.3.

The U-Net, R2U-Net, and AttU-Net implementations followed the original architectural specifications as described in their respective publications [73, 74, 75]. The convolutional autoencoder architecture was designed to mirror the encoder layout of U-Net. Specifically, the CAE encoder consisted of four convolutional layers with $3 \times 3$ kernels, each followed by a ReLU activation and $2 \times 2$ max pooling. These layers comprised 64, 128, 256, and 512 filters, respectively. A final convolutional layer with 1,024 filters formed the bottleneck representation, which was then passed to a decoder mirroring the encoder's structure. For all architectures, the extracted image features were flattened prior to clustering.

**To extract feature vectors from narrative diagnosis texts**, three commonly used text embedding methods were evaluated: Bag of Words (BoW) [76], Term Frequency–Inverse Document Frequency (TF-IDF) [77], and Doc2Vec [78]. These methods were selected based on their popularity and proven effectiveness in prior surveys focused on text representation in clinical contexts [79, 80, 81, 82]. Each approach was applied to the corpus of narrative diagnoses present in the training subset of the dataset. Hyperparameter value ranges for all tested methods are listed in Table 2.4. Each embedding method has its own benefits and caveats. BoW offers a simple and computationally efficient representation but fails to account for word frequency or contextual semantics. TF-IDF improves on BoW by weighting terms according to their relative frequency, thus reducing the influence of commonly used words, though it still lacks a mechanism for capturing word order or context. Doc2Vec, a neural embedding method, addresses these limitations by learning fixed-length vector representations of entire documents (in this case, diagnostic texts). Two variants of Doc2Vec were tested: Paragraph Vector–Distributed Memory (PV-DM), which predicts a target word based on its context and the paragraph vector; and Paragraph Vector–Distributed Bag of Words (PV-DBOW), which predicts randomly

sampled words in the paragraph using only the paragraph vector. The dimensionality of the output embeddings was controlled via the embedding size hyperparameter. In PV-DM, the input includes both the paragraph vector and a context window of words, while PV-DBOW relies solely on the paragraph vector.

### 2.4.2. Grouping

Clustering was conducted independently on each of the three data sources (DICOM tags, images, and textual diagnosis). In each case, the raw input data were first pre-processed and transformed using the feature extraction methods described previously. The resulting feature embeddings served as input to clustering algorithms: k-means [86] and k-medoids [87]. The difference between these two algorithms is that k-means uses the mean of points in a cluster as the centroid, which may or may not correspond to an actual data point and makes it sensitive to outliers. In contrast, k-medoids selects a medoid, an actual data point with the lowest total distance to others in the cluster, making it more robust to outliers. An intuitive analogy to conceptually comparing k-means and k-medoids is to consider the difference of *mean* versus *median* value. K-means finds the *mean* value (centroid) in multidimensional space, while the goal of k-medoids is to find the *median* value (medoid). To compute centroids, k-means always uses the Euclidean (L2) distance, while k-medoids can compute medoids based on different distance metrics. In this case, for k-medoids, Euclidean and cosine distance metrics were tested to account for potential differences in the structure of the feature space.

Clustering was performed for $\kappa \in \{5, 10, 15, 20, 25, 30, 40, 50, 75, 100, 150\}$; while higher values of $\kappa$ were initially considered, they were ultimately excluded due to issues encountered during experimentation, including a large number of empty clusters, significant inter-cluster overlap, or other signs of overfitting. As clustering outcomes can be sensitive to the initialisation of centroids or medoids, robustness was assessed by performing 11 independent runs of the best-performing configurations. No statistically significant variation was observed across these runs, suggesting stability of the clustering solutions with respect to initialisation.

**Table 2.3:** Explored hyperparameter value ranges for DICOM tags, image and diagnosis feature extraction. The values were originally taken from related work, and then refined empirically [41].

| **Data Source** | **Feature extractor(s)** | **Hyperparameter** | **Tested values** |
|---|---|---|---|
| DICOM tags | AE | Learning rate | $10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$ |
| | | Bottleneck size (embedding size) | 32, 64, 50, 75, 100 |
| | | Layer sizes | $1^{st}$: 300, 250, 200, 512, 256<br>$2^{nd}$: 200, 150, 125, 100, 256, 128<br>$3^{rd}$: 150, 125, 100, 75, 128, 64 |
| | PCA | Solver | LAPACK [83], ARPACK [84], randomised [85] |
| | | Number of components (embedding size) | 32, 64, 50, 75, 100 |
| Images | CAE, U-Net, AttU-Net, R2U-Net | Learning rate | $10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}$ |
| | | PCA solver | [Not applied], LAPACK, ARPACK, randomised |
| | | PCA number of components (embedding size) | [Not applied], 100, 500, 1000, 10000 |
| Diagnoses | BOW, TF-IDF, PV-DM, PV-DBOW [76] | Minimum word frequency | 5, 10, 50, 100, 500, 1000, 2000, 3500, 5000, 10000 |
| | PV-DM, PV-DBOW | Embedding size | 10, 25, 50, 100, 250, 500, 1000 |
| | | Window size | 5, 7, 10 |
| | | Number of epochs | 25, 50, 100 |

**Evaluation metrics**

To evaluate the quality of the resulting cluster assignments, particular emphasis was placed on assessing cluster homogeneity with respect to imaging modality and body region. Accordingly, homogeneity score (HS) and normalised mutual information (NMI) were computed based on the `Modality` and `BodyPartExamined` DICOM tags. Both metrics range from 0.00 to 1.00, where higher values indicate greater alignment between the cluster assignments and the reference labels. Let $y_M$ denote the ground truth imaging modality, $\hat{y}$ the predicted cluster label, $I(y_M, \hat{y})$ the mutual information between the two, and $H(y_M)$ and $H(\hat{y})$ their respective entropies. The NMI score with respect to modality, denoted $NMI_M$, is defined as [88]:

$$NMI_M = \frac{2 \cdot I(y_M, \hat{y})}{H(y_M) + H(\hat{y})}. \tag{2.6}$$

$I(y_M, \hat{y})$ can be calculated as $I(y_M, \hat{y}) = H(y_M) - H(y_M|\hat{y})$, where $H(y_M|\hat{y})$ is the conditional entropy. HS regarding modality ($HS_M$) can be calculated as:

$$HS_M = 1 - \frac{H(y_M|\hat{y})}{H(y_M)}. \tag{2.7}$$

It is important to note that the denominator in the homogeneity score calculation for modality ($HS_M$) is guaranteed to be non-zero, as the label distribution in the observed subset is not perfectly balanced (i.e. the subset is not monotonically pure). The same procedure applies when computing $NMI$ and $HS$ for the body part examined ($NMI_B$ and $HS_B$), respectively. In this case, $y_M$ is replaced with $y_B$, corresponding to the `BodyPartExamined` tag. The predicted cluster assignments $\hat{y}$ remain unchanged across all calculations.

Finally, the overall clustering quality was summarised by computing the harmonic mean of the four evaluation metrics: $HS_B$, $HS_M$, $NMI_B$, and $NMI_M$. This aggregate score (denoted as $S$), provides a single, comprehensive measure of clustering performance that accounts for both modality- and anatomy-based homogeneity and mutual information.

In addition to cluster homogeneity with respect to imaging modality and anatomical region, high-quality clustering should also reflect semantic and visual coherence across

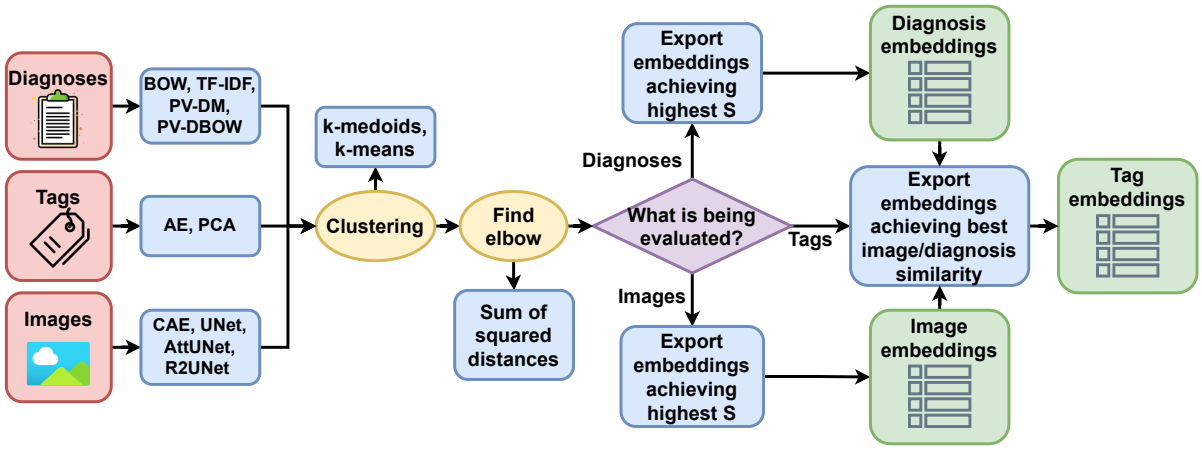



**Figure 2.9:** Evaluation pipeline for all three data sources [41].

other modalities – specifically, the image content and associated narrative diagnoses. It is expected that data points grouped within the same cluster will exhibit visual similarity in terms of image structure and semantic similarity in their textual diagnoses. To evaluate this, cosine distances were computed for both image and diagnosis embeddings. The procedure for evaluating intra-cluster image similarity is as follows. Let $k$ data points be assigned to a cluster indexed by $c$, where $0 \leq c < \kappa$. For each distinct pair of data points $(i, j)$ such that $i \neq j$ and both $i, j \in c$, let $x_I^{(i)}$ and $x_I^{(j)}$ denote the corresponding image instances, and $f(x_I^{(i)})$, $f(x_I^{(j)})$ their respective feature embeddings. The cosine distance $d$ between the two embeddings is given by:

$$d(x_I^{(i)}, x_I^{(j)}) = 1 - \frac{f(x_I^{(i)})^T \cdot f(x_I^{(j)})}{\|f(x_I^{(i)})\| \cdot \|f(x_I^{(j)})\|}. \tag{2.8}$$

The possible number of pairs in cluster $c$ is $u^{(c)} = \frac{k}{2}(k-1)$. To find the dissimilarity of images in cluster $D_I^{(c)}$, calculate the mean cosine distance of all pairs from cluster $c$:

$$D_I^{(c)} = \frac{1}{u^{(c)}} \sum_{i=0}^{k-1} \sum_{j=i+1}^{k-1} d(x_I^{(i)}, x_I^{(j)}), \tag{2.9}$$

and finally, overall image similarities across all clusters were computed as $D_I = \frac{1}{\kappa} \sum_{c=0}^{\kappa-1} D_I^{(c)}$.

The same process was applied to get diagnoses similarities $D_D$ from diagnoses embeddings. Ideally, $D_I$ and $D_D$ should be close to 0, i.e. the distances between embeddings in the same cluster should be as small as possible, which indicates that diagnoses / images in the same cluster exhibit clear similarities.

### Evaluation process

To identify the most informative data source embeddings, clustering performance was compared across all feature extractors and data modalities using the validation set. The overall evaluation workflow is illustrated in Figure 2.9. To determine the optimal number of clusters for each data source, the elbow method [89] was applied by analysing the sum of squared distances from each data point to its nearest cluster centre. Elbow points were detected using the Kneedle algorithm proposed by Satopaa et al. [90]. Having too few clusters could result in a heterogeneous grouping, while having too many clusters might lead to groups that are homogeneous but show evidence of incompleteness.

Different sources of data were evaluated using different metrics, as illustrated in Figure 2.9. The best feature extractors for images and textual diagnoses were chosen based on the highest $S$ score. These feature extractors were later used to evaluate the efficiency of DICOM tag clustering, i.e. to compute image ($D_I$) and diagnosis ($D_D$) (dis)similarities. Finally, to rank the efficiency of DICOM tag feature extraction models, $D_{score}$ was calculated as the harmonic mean of $D_I$ and $D_D$. The primary objective is to create clusters that exhibit the highest degree of data similarity. Therefore, the model obtaining the lowest $D_{score}$ value at the elbow would be selected as the best DICOM tag feature extraction model.

### Feature fusion

After selecting the optimal feature extractor for each data source, the resulting embeddings were fused using three distinct strategies: (i) direct concatenation of raw embeddings, (ii) concatenation of cluster-space distances, and (iii) concatenation of soft cluster probability assignments. In each of the approaches, the resulting vector of a single data point $i$ was flat, and in the format of $f(x^{(i)}) = \left[f(x_D^{(i)}), f(x_T^{(i)}), f(x_I^{(i)})\right]^T$, where $f(x_D^{(i)})$ is the diagnosis embedding, $f(x_T^{(i)})$ is the DICOM tags embedding, and $f(x_I^{(i)})$ is the image embedding.

The first fusion strategy involved straightforward concatenation of the raw embeddings from the three sources into a single feature vector. This simple approach was also used in related work, such as [91], where embeddings from tabular data and clinical text were combined to form a unified representation.

The second strategy, henceforth referred to as *clusterdists*, used distances in cluster space as feature representations. During clustering, distances to each of the cluster centres are computed, and then the point is assigned to the nearest cluster. Embeddings carrying similar information should also have similar distances to each of the cluster centres. Hence, instead of using the extracted embeddings, the computed distances to each cluster centre were used and subsequently concatenated together. All distances were normalised to fit the range $[0.00, 1.00]$ before concatenation.

The third strategy, denoted as *clusterprobs*, extended the *clusterdists* approach by converting distances to soft cluster assignment probabilities using a softmax function. For a given data point $i$, the probability of belonging to cluster $c$ is computed as:

$$p_c^{(i)} = \frac{e^{-d_c^{(i)}}}{\sum_{j=1}^{\kappa} e^{-d_j^{(i)}}}, \tag{2.10}$$

where $\kappa$ is the number of clusters, and $d_c$ and $d_j$ are distances to $c$-th and $j$-th cluster of the respective source embeddings.

Alternative fusion strategies were considered, such as those involving attention-based joint encoders or contrastive pretraining frameworks as proposed by Radford et al. [92]. However, these methods were ultimately excluded due to the computational demands imposed by the scale of the dataset and hardware limitations. At the time, the available hardware included a server with 2× AMD EPYC 7702 64-Core Processor and 1 terabyte of RAM memory, but with limited GPU power. While two Gigabyte GeForce RTX 3090 GPUs were present, GPU-based workloads were limited due to stability and memory allocation issues. As a result, the computational decisions in the experimental setup were made to optimise performance within the available resources.

### 2.4.3. Results

Each of the described feature extractors were tested across multiple hyperparameter values with an extensive ablation study. The results are presented here, with the first subsection presenting the performance of each feature extractor for each data type. This is followed by a subsection detailing a thorough ablation study, where all possible combinations of feature extractors were tested and evaluated, to compare how the features of each data source contributed towards the final pseudo-labels.

### Optimal embeddings

A total of ten models were trained for feature extraction across the three data modalities: four models for narrative diagnoses (TF-IDF, BoW, PV-DM, PV-DBOW), four models for image data (CAE, U-Net, AttU-Net, R2U-Net), and two models for DICOM metadata (AE and PCA). The embeddings produced by each extractor were evaluated across all tested hyperparameter configurations and clustering strategies. The optimal hyperparameter values for each of the four model categories are summarised in Table 2.4, while the clustering performance of the best-performing models is reported in Table 2.5.

Results show that CAE was the best-performing image feature extractor. Compared to U-Net, AttU-Net, and R2U-Net, CAE achieved superior cluster homogeneity with respect to both imaging modality and examined body part, yielding the highest scores across all four evaluation metrics ($HS_M$, $HS_B$, $NMI_M$, and $NMI_B$) on the validation set. Consequently, it also achieved the highest aggregate score $S$. Among the feature extractors for narrative diagnoses, the PV-DBOW model produced the best overall performance. It attained the highest $HS_M$, $HS_B$, and $NMI_M$ scores, while its $NMI_B$ was slightly lower than that of TF-IDF. Nonetheless, the aggregate $S$ score confirmed PV-DBOW as the most effective text embedding method in this context.

To calculate image ($D_I$) and diagnoses ($D_D$) distances and the corresponding $D_{score}$ for DICOM tag evaluation, the best-performing feature extractors from images and diagnoses were used, i.e. CAE and PV-DBOW. As it can be seen in Table 2.5, the best image and diagnosis similarity on the validation subset was achieved using AE.

A more detailed analysis of the clustering performance for the best models is provided in Figure 2.10. Specifically, Figure 2.10a shows CAE performance for image embeddings, Figure 2.10b shows PV-DBOW results for textual diagnoses, and Figure 2.10c shows the AE performance for DICOM metadata. From the shown metrics, it is evident that the models perform nearly the same on the validation and test set.

### Source fusion ablation study

Based on the obtained results, the selected models for feature fusion were: **AE** for DICOM tags, **CAE** for images and **PV-DBOW** for narrative diagnoses. The next goal was to thoroughly investigate the relationship between clustering results and embedding

**Table 2.4:** Hyperparameter values of best performing feature extraction models (emphasised) for each data source independently and all feature extractors for respective sources covered by our experiments (Table 2.3) [41].

| Data source | Model name | Hyperparameters | Embedding size | Algorithm | Cluster metric | Number of clusters |
|---|---|---|---|---|---|---|
| Images | **CAE** | Learning rate: $10^{-6}$<br>PCA solver: randomised | 500 | k-means | Euclidean | 40 |
| | U-Net | Learning rate: $10^{-6}$<br>PCA solver: randomised | 10,000 | k-means | Euclidean | 30 |
| | AttU-Net | Learning rate: $10^{-6}$<br>PCA not applied | 65,536 | k-medoids | cosine | 30 |
| | R2U-Net | Learning rate: $10^{-6}$<br>PCA solver: randomised | 5,000 | k-medoids | cosine | 40 |
| Diagnoses | TF-IDF | Minimum word frequency: 1,000 | 825 | k-means | Euclidean | 30 |
| | BOW | Minimum word frequency: 10 | 7,951 | k-means | Euclidean | 30 |
| | PV-DM | Minimum word frequency: 100<br>Window size: 5<br>Number of epochs: 50 | 10 | k-means | Euclidean | 20 |
| | **PV-DBOW** | Minimum word frequency: 50<br>Window size: 7<br>Number of epochs: 50 | 1,000 | k-means | Euclidean | 20 |
| DICOM tags | **AE** | $512 \rightarrow 200 \rightarrow 125 \rightarrow 32$<br>Learning rate: $10^{-2}$ | 32 | k-medoids | cosine | 25 |
| | PCA | Solver: ARPACK | 50 | k-medoids | Euclidean | 40 |

**Table 2.5:** Results for each best performing feature extraction model for images, diagnoses and tag clustering, computed on the validation subset. Best results are emphasised. ± sign delimits the mean from the standard deviation [41].

| Data source | Model name | $NMI_B$ | $NMI_M$ | $HS_B$ | $HS_M$ | $S$ | $D_D[\cdot 10^{-2}]$ | $D_I[\cdot 10^{-2}]$ | $D_{score}$ |
|---|---|---|---|---|---|---|---|---|---|
| Images | **CAE** | **0.430** | **0.459** | **0.538** | **0.744** | **0.519** | - | - | - |
| | U-Net | 0.401 | 0.450 | 0.467 | 0.673 | 0.479 | - | - | - |
| | AttU-Net | 0.285 | 0.348 | 0.328 | 0.512 | 0.351 | - | - | - |
| | R2U-Net | 0.186 | 0.206 | 0.185 | 0.255 | 0.205 | - | - | - |
| Diagnoses | TF-IDF | **0.603** | 0.493 | 0.623 | 0.716 | 0.598 | - | - | - |
| | BOW | 0.529 | 0.508 | 0.576 | 0.788 | 0.583 | - | - | - |
| | PV-DM | 0.544 | 0.467 | 0.570 | 0.689 | 0.557 | - | - | - |
| | **PV-DBOW** | 0.592 | **0.569** | **0.634** | **0.77** | **0.633** | - | - | - |
| DICOM tags | **AE** | - | - | - | - | - | **3.553 ± 2.11** | **3.051 ± 1.01** | **3.283** |
| | PCA | - | - | - | - | - | 3.578 ± 2.019 | 3.248 ± 1.171 | 3.405 |

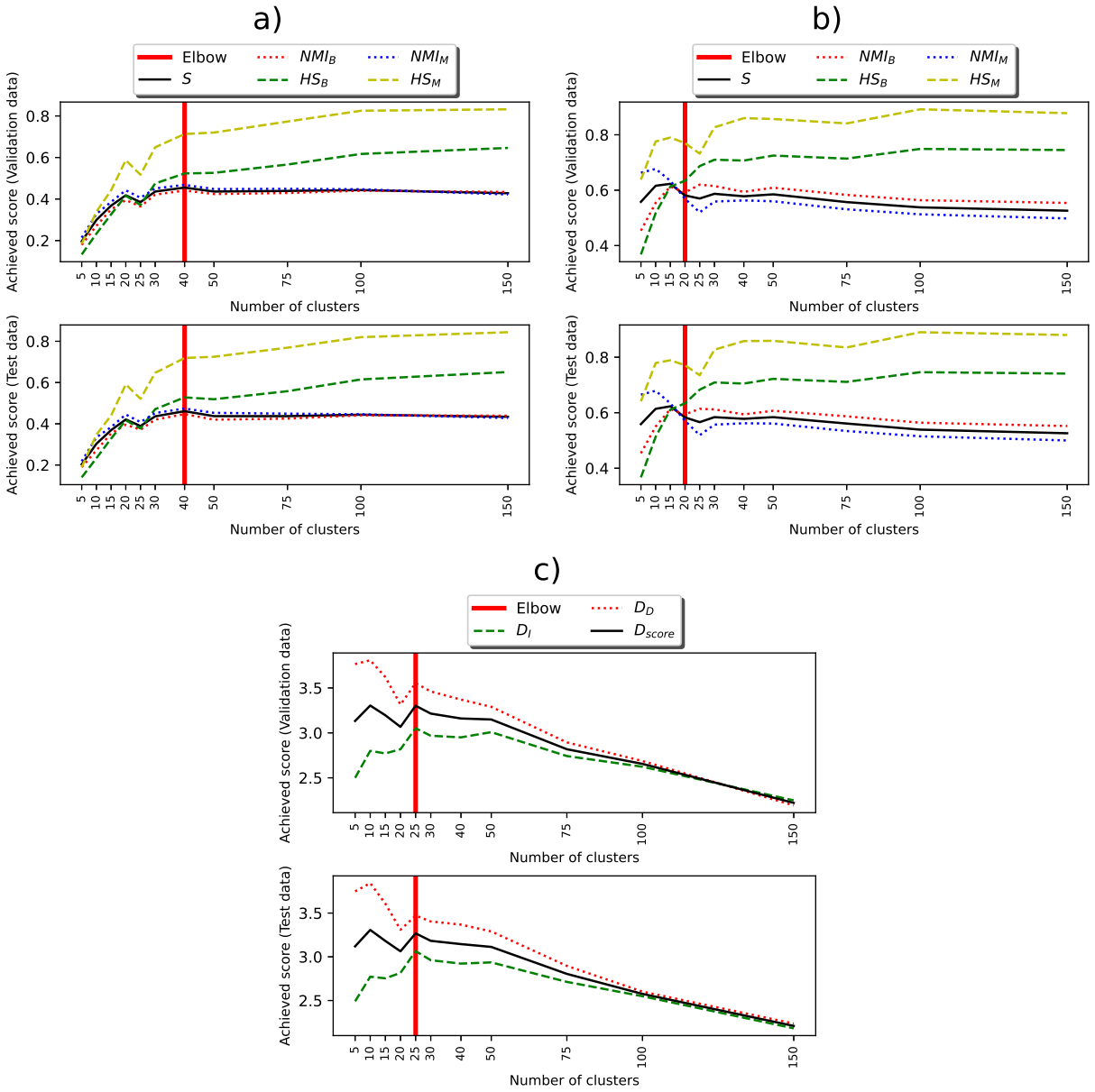



**Figure 2.10:** Diagrams showing individual evaluation metrics values on validation (top) and test (bottom) subsets, when clustering optimal image embeddings (CAE, subfigure a), optimal diagnoses embeddings (PV-DBOW, subfigure b) and DICOM tag embeddings (AE, subfigure c) [41].

**Table 2.6:** Hyperparameter values of each best performing model obtained by fusing individual source embeddings. Model performance is shown in Table 2.7). In this table, *AE* is used to describe DICOM tag embeddings, *CAE* image embeddings, and *PV-DBOW* diagnosis embeddings [41].

| Model name | Combine Method | Embedding size | Algorithm | Clustering metric | Num clusters |
|---|---|---|---|---|---|
| [AE]-[CAE] | clusterdists | 65 | k-means | Euclidean | 30 |
| | clusterprobs | 65 | k-medoids | cosine | 30 |
| | embeddings | 532 | k-means | Euclidean | 75 |
| [PV-DBOW]-[AE] | clusterdists | 45 | k-means | Euclidean | 30 |
| | clusterprobs | 45 | k-medoids | Euclidean | 30 |
| | embeddings | 1032 | k-means | Euclidean | 40 |
| [PV-DBOW]-[CAE] | clusterdists | 60 | k-medoids | cosine | 30 |
| | clusterprobs | 60 | k-medoids | Euclidean | 40 |
| | embeddings | 1500 | k-means | Euclidean | 40 |
| [PV-DBOW]-[AE]-[CAE] | clusterdists | 85 | k-medoids | Euclidean | 40 |
| | clusterprobs | 85 | k-medoids | cosine | 40 |
| | embeddings | 1532 | k-means | Euclidean | 50 |

**Table 2.7:** Results for each best-performing model for all combinations of data sources, computed on the validation subset. Best results are emphasised for each specific metric utilised. ± sign delimits the mean from the standard deviation. In this table, *AE* is used to describe DICOM tag embeddings, *CAE* image embeddings, and *PV-DBOW* diagnosis embeddings [41].

| Model name | Combine Method | $NMI_B$ | $NMI_M$ | $HS_B$ | $HS_M$ | $S$ | $D_D[\cdot 10^{-2}]$ | $D_I[\cdot 10^{-2}]$ | $D_{score}$ |
|---|---|---|---|---|---|---|---|---|---|
| [AE]-[CAE] | clusterdists | 0.486 | 0.652 | 0.574 | 0.99 | 0.631 | 3.599 ± 1.976 | 2.663 ± 1.129 | 3.061 |
| | clusterprobs | 0.503 | 0.68 | 0.578 | 0.999 | 0.647 | 3.564 ± 2.061 | 2.759 ± 1.101 | 3.11 |
| | embeddings | 0.462 | 0.534 | 0.613 | 0.928 | 0.593 | 3.232 ± 1.787 | 2.268 ± 0.976 | 2.666 |
| [PV-DBOW]-[AE] | clusterdists | 0.46 | 0.675 | 0.532 | 0.999 | 0.612 | **2.894 ± 2.356** | 2.717 ± 1.039 | 2.802 |
| | clusterprobs | 0.516 | **0.697** | 0.581 | **1.0** | 0.656 | 3.385 ± 2.205 | 2.802 ± 1.045 | 3.066 |
| | embeddings | 0.609 | 0.563 | 0.712 | 0.845 | **0.666** | 3.228 ± 1.891 | 2.179 ± 0.863 | 2.602 |
| [PV-DBOW]-[CAE] | clusterdists | 0.373 | 0.429 | 0.44 | 0.651 | 0.453 | 2.927 ± 2.038 | 2.457 ± 0.965 | 2.671 |
| | clusterprobs | 0.425 | 0.453 | 0.529 | 0.732 | 0.512 | 3.269 ± 1.878 | 2.272 ± 1.031 | 2.681 |
| | embeddings | **0.611** | 0.548 | **0.713** | 0.819 | 0.657 | 3.234 ± 1.957 | 2.163 ± 0.879 | 2.592 |
| **[PV-DBOW]-[AE]-[CAE]** | clusterdists | 0.472 | 0.616 | 0.583 | 0.986 | 0.618 | 3.362 ± 2.088 | 2.587 ± 1.104 | 2.924 |
| | clusterprobs | 0.497 | 0.639 | 0.604 | **1.0** | 0.642 | 3.483 ± 1.996 | 2.779 ± 1.134 | 3.091 |
| | **embeddings** | 0.587 | 0.57 | 0.707 | 0.885 | **0.666** | 2.982 ± 1.688 | **2.012 ± 0.783** | **2.403** |

sources included in clustering. To this end, a comprehensive hyperparameter analysis was conducted using the validation set, leading to the selection of the optimal hyperparameter values. These values are shown in Table 2.6, while the corresponding performance on the validation set is presented in Table 2.7.The primary criterion for selecting the optimal hyperparameters was the aggregate score $S$, while the metric $D_{score}$ was also considered when the $S$ score alone was insufficient to distinguish between the best results.

**DICOM tags and images** ([AE]-[CAE]): When examining the combination of DICOM tags and images, the results in Table 2.7 compared to those in Table 2.5 indicate that combining DICOM tags with image features leads to an improvement in the $D_{score}$ relative to using AE alone. Furthermore, all three fusion methods (*embeddings*, *clusterdists*, and *clusterprobs*) achieved higher $S$ scores compared to using image features alone, demonstrating improved modality and anatomical region homogeneity.

**Diagnoses and DICOM tags** ([PV-DBOW]-[AE]): Following a similar pattern to [AE]-[CAE], combining DICOM tags with diagnoses results in an improvement in the $D_{score}$ compared to using DICOM tags alone. When applying the *embeddings* combination method, the $S$ score is higher than that obtained using diagnoses alone (and represents the highest overall score on the validation subset), with a noticable improvement, particularly in $HS_B$ and $HS_M$. Moreover, when employing the *clusterdists* and *clusterprobs* methods, the combined approach shows a notable increase in modality homogeneity compared to using diagnosis embeddings alone. However, a trade-off is observed in terms of examined body part homogeneity, as $NMI_B$ and $HS_B$ are lower in the *clusterdists* and *clusterprobs* approaches compared to the $HS_B$ and $NMI_B$ obtained by diagnoses alone.

**Diagnoses and images** ([PV-DBOW]-[CAE]): Combining images with diagnoses, particularly using the *embeddings* method, results in the most accurate grouping by examined body part, yielding the highest overall $NMI_B$ and $HS_B$. The $S$ score obtained through this combination is higher than the $S$ scores achieved when using images and diagnoses independently.

**Diagnoses, DICOM tags, and images** ([PV-DBOW]-[AE]-[CAE]): Finally, when all three data sources are combined using the *embeddings* method, the highest overall $S$ score is achieved. This score matches the $S$ score obtained using the [PV-DBOW]-[AE] combination (diagnoses and DICOM tags) with the *embeddings* method; however, there is a noticeable difference in the $D_{score}$ between them. Alternatively, when all three data

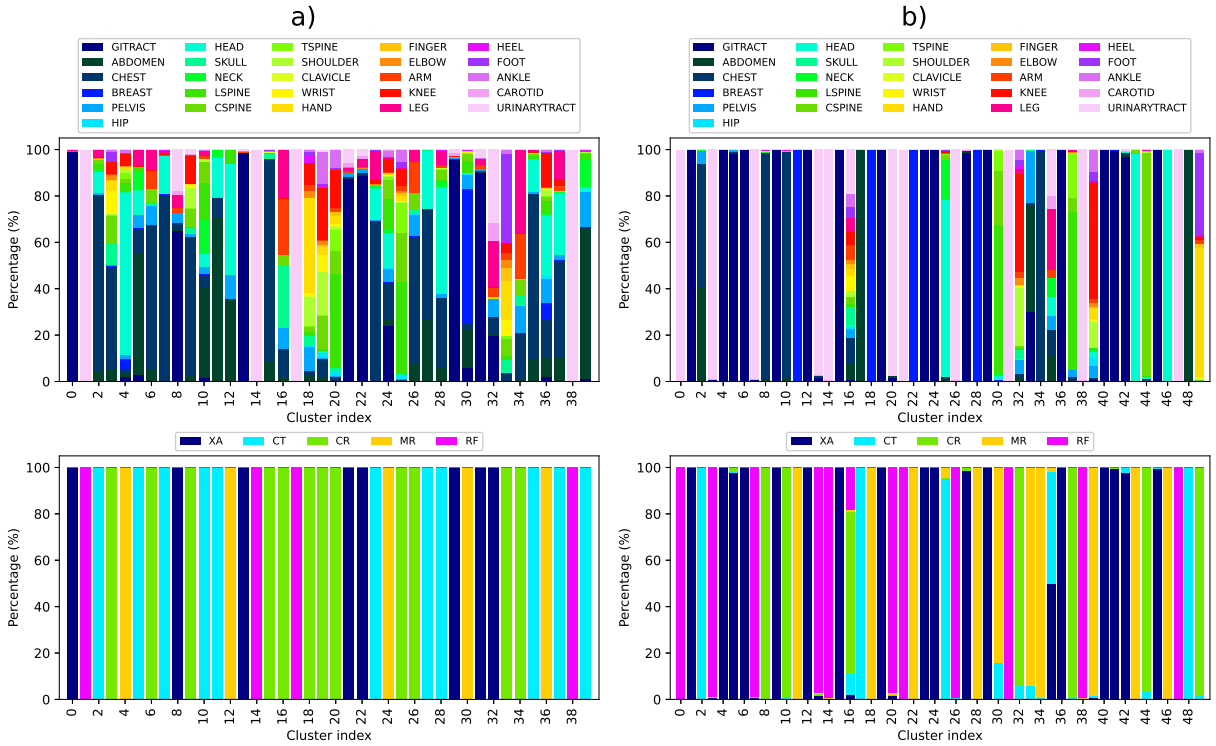



**Figure 2.11:** Grouping quality regarding modality and examined body part, when grouping by [PV-DBOW]-[AE]-[CAE] using *clusterprobs* (subfigure a) and *embeddings* (subfigure b) combine methods [41].

sources are combined using the *clusterprobs* method, a perfect $HS_M$ score is obtained. Figure 2.11a and Figure 2.11b illustrate the variation in grouping quality achieved through these two combination methods (*clusterprobs* versus *embeddings*). In these figures, each bar represents the mixture ratio within a specific cluster, where the images on top show how homogeneous the clusters are when observing the body part (i.e. how mixed the clusters are with regard to anatomical region), while the bottom images show the different modalities in each cluster (i.e. how mixed the clusters are with regard to imaging modality).

Finally, the performance of all individual data sources, as well as all feature combinations and combination methods on the test subset, is presented in Table 2.8.

### 2.4.4. Discussion

As shown in Table 2.8, among the three data sources, image clustering using CAE exhibited the poorest performance with respect to both modality and examined body part homogeneity. Upon visual inspection of the resulting groups, it was observed that

**Table 2.8:** Clustering results on the test subset, when using the best performing models from all data sources and all three feature fusion approaches. ± sign delimits the mean from the standard deviation [41].

| **Model name** | **Combine Method** | $NMI_B$ | $NMI_M$ | $HS_B$ | $HS_M$ | $S$ | $D_D[\cdot 10^{-2}]$ | $D_I[\cdot 10^{-2}]$ | $D_{score}$ |
|---|---|---|---|---|---|---|---|---|---|
| AE (DICOM Tags) | - | - | - | - | - | - | 3.471 ± 2.059 | 3.065 ± 1.107 | 3.255 |
| CAE (Images) | - | 0.447 | 0.474 | 0.528 | 0.719 | 0.524 | - | - | - |
| PV-DBOW (Diagnoses) | - | 0.594 | 0.571 | 0.634 | 0.771 | 0.634 | - | - | - |
| [AE]-[CAE] | clusterdists | 0.494 | 0.652 | 0.584 | 0.992 | 0.637 | 3.592 ± 1.941 | 2.654 ± 1.146 | 3.053 |
| | clusterprobs | 0.507 | 0.678 | 0.584 | **1.0** | 0.65 | 3.546 ± 2.026 | 2.769 ± 1.109 | 3.109 |
| | embeddings | 0.465 | 0.539 | 0.615 | 0.933 | 0.597 | 3.22 ± 1.711 | 2.256 ± 0.983 | 2.653 |
| [PV-DBOW]-[AE] | clusterdists | 0.462 | 0.671 | 0.536 | 0.999 | 0.614 | **2.906 ± 2.358** | 2.697 ± 1.058 | 2.797 |
| | clusterprobs | 0.523 | **0.69** | 0.593 | **1.0** | 0.662 | 3.364 ± 2.172 | 2.801 ± 1.09 | 3.057 |
| | embeddings | 0.605 | 0.561 | 0.711 | 0.847 | 0.664 | 3.172 ± 1.815 | 2.179 ± 0.921 | 2.584 |
| [PV-DBOW]-[CAE] | clusterdists | 0.368 | 0.424 | 0.434 | 0.642 | 0.447 | 2.913 ± 1.991 | 2.442 ± 0.967 | 2.657 |
| | clusterprobs | 0.427 | 0.455 | 0.532 | 0.736 | 0.514 | 3.279 ± 1.85 | 2.269 ± 1.039 | 2.682 |
| | embeddings | **0.611** | 0.546 | **0.712** | 0.816 | 0.656 | 3.208 ± 1.92 | 2.182 ± 0.93 | 2.598 |
| **[PV-DBOW]-[AE]-[CAE]** | clusterdists | 0.478 | 0.617 | 0.591 | 0.989 | 0.623 | 3.354 ± 2.096 | 2.571 ± 1.1 | 2.911 |
| | clusterprobs | 0.499 | 0.637 | 0.606 | **1.0** | 0.643 | 3.47 ± 1.94 | 2.787 ± 1.173 | 3.091 |
| | **embeddings** | 0.587 | 0.571 | 0.705 | 0.885 | **0.666** | 3.024 ± 1.591 | **2.011 ± 0.828** | **2.416** |

although images within the same cluster can appear visually similar, they often depict different body regions captured by the same modality or vice-versa, the same body region captured using different modalities. Furthermore, variations in windowing parameters can significantly affect the visual appearance of images, contributing to inconsistencies in clustering. An illustrative example is shown in cluster 2 (Figure 2.12), where all images represent parts of the abdomen but are clearly visually different. These suggests that images alone do not provide sufficient information for reliable semantic clustering. However, as indicated in Table 2.8, integrating image data into the clustering process reduces the $D_{score}$ and enhances the visual similarity within clusters. Thus, while images alone may lack the necessary semantic information for optimal grouping, their inclusion contributes to improved cluster representation when combined with other data sources.

Diagnoses (PV-DBOW) demonstrated good performance in terms of anatomical region grouping. This outcome can be attributed to the fact that the content diagnoses often explicitly references the anatomical region being examined, typically by describing pathologies (illnesses or injuries) or procedures affecting a specific body part. As shown in Table 2.8, when diagnoses are integrated with other data sources, the quality of grouping by examined body part shows a noticeable improvement. Therefore, it can be inferred that the inclusion of narrative diagnoses enhances anatomical region homogeneity within clusters.

Wherever DICOM tags (AE) were used, a noticeable improvement in modality homogeneity was observed in Table 2.8. When combined with DICOM tags, image embeddings demonstrated superior $NMI_M$ and $HS_M$ scores compared to using images alone. A similar trend was observed for diagnoses, where the inclusion of DICOM tags resulted in an increase in $HS_M$. These findings indicate that DICOM tags contribute to improved modality homogeneity. This can be attributed to the structure of the DICOM standard [42], as the DICOM tags frequently contain modality-specific values.

Three different feature fusion approaches were evaluated, each demonstrating satisfactory performance. Two alternative methods for feature fusion, namely *clusterprobs* and *clusterdists*, were introduced and tested. Compared to the conventional *embeddings* approach, both *clusterprobs* and *clusterdists* exhibited a tendency to prioritise high modality homogeneity. In particular, three model configurations presented in Table 2.8 ([PV-DBOW]-[AE]-[CAE] *clusterprobs*, [AE]-[CAE] *clusterprobs*, and [PV-DBOW]-[AE]

*clusterprobs*) achieved perfect scores for $HS_M$. Nevertheless, the *embeddings* approach consistently outperformed in terms of anatomical region homogeneity. This difference is especially evident in Figure 2.11a and Figure 2.11b, where the former prioritises modality homogeneity, while the latter demonstrates better grouping by anatomical region.

As is visible in Table 2.8, some models achieved comparable $S$ scores. However, among these models, there is a noticeable difference in $D_{score}$, where a lower $D_{score}$ indicates that the clusters are more visually homogeneous and contain diagnoses with higher semantic similarity (which is considered favourable). Therefore, in this context, the approach that combines all three data sources using the *embeddings* method can be regarded as the most effective for achieving optimal grouping with respect to modality, examined body part, and the similarity of both images and diagnoses.

## 2.5. Pseudo-labels and the annotated dataset

From the original dataset (described in section *2.3.*), a total of 1,337,926 data points that satisfied the specified criteria related to image quality, diagnosis completeness, and DICOM tag consistency were extracted. The selected labelling algorithm, illustrated in Figure 2.13, was used to cluster this expanded set of data points into 50 groups, whose sizes are shown in 2.14. As it can be seen in Figure 2.14, the resulting clusters exhibited variable sizes, with the largest containing 341,083 data points, and the smallest consisting of only 6 data points. Small clusters of this nature can be considered as outliers or anomalies, indicating data points that do not fit into the patterns present in larger groups.

The overall quality of the obtained clusters is shown in Figure 2.15, which demonstrates that the grouping quality remains consistent with that observed in Figure 2.11b. This consistency indicates that incorporating previously unseen data did not significantly affect the quality of the clusters. Groups which were heterogeneous on the smaller set (16, 32, 35 and 39) remained the same after labelling the larger set, and the same applies to homogeneous clusters such as 2, 3, 8, 25, 43 and 44. Random instances of images from these (and other) clusters were provided in Figure 2.12.

Next, considering the quality of the obtained groups presented in Figure 2.15, it is evident that most clusters exhibit a high degree of homogeneity with respect to imaging modality. In contrast, anatomical region homogeneity reveals that neighbouring anatom-

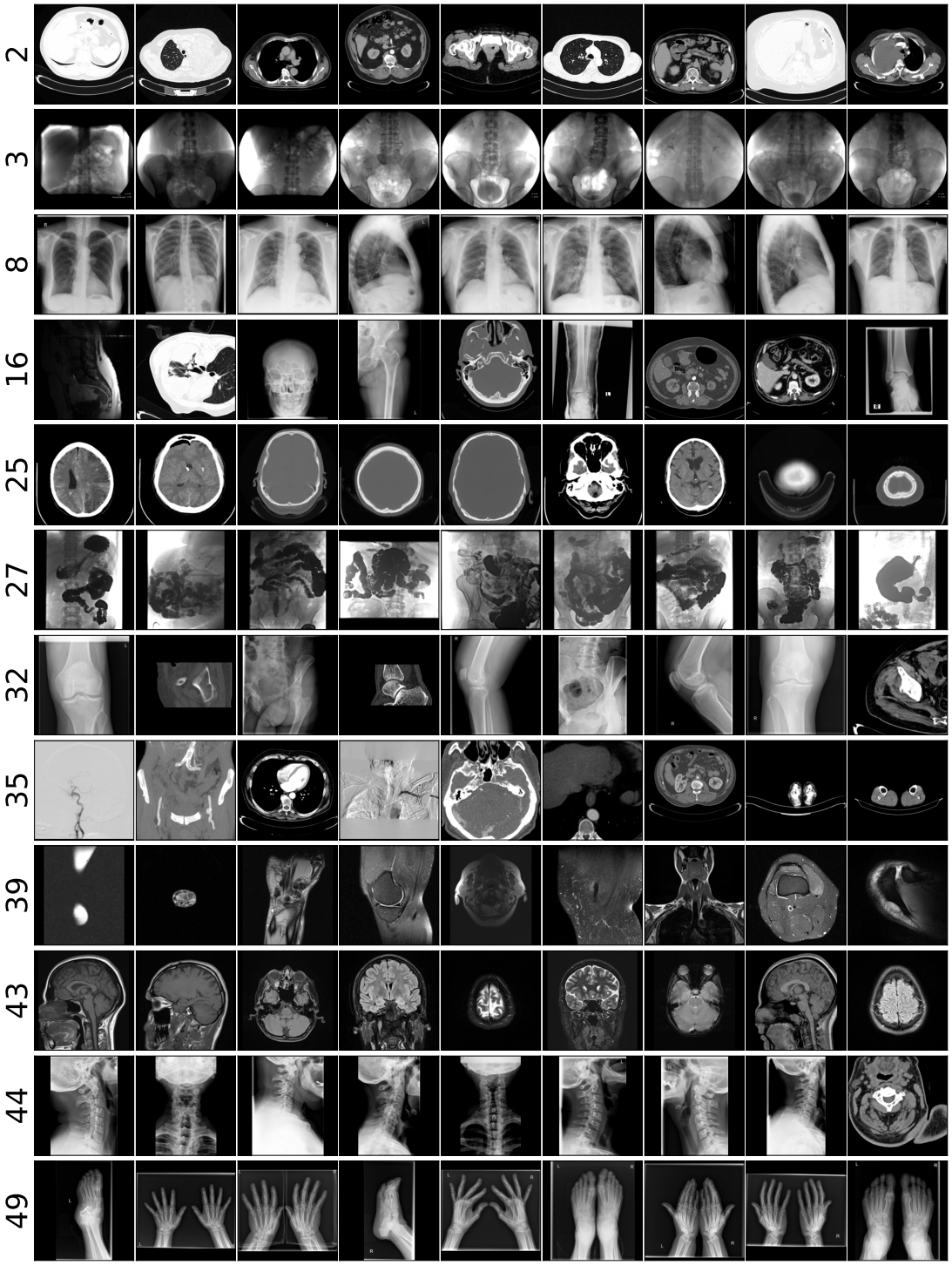



**Figure 2.12:** Randomly sampled images from twelve selected clusters. Cluster indices are indicated to the left of each row [41].

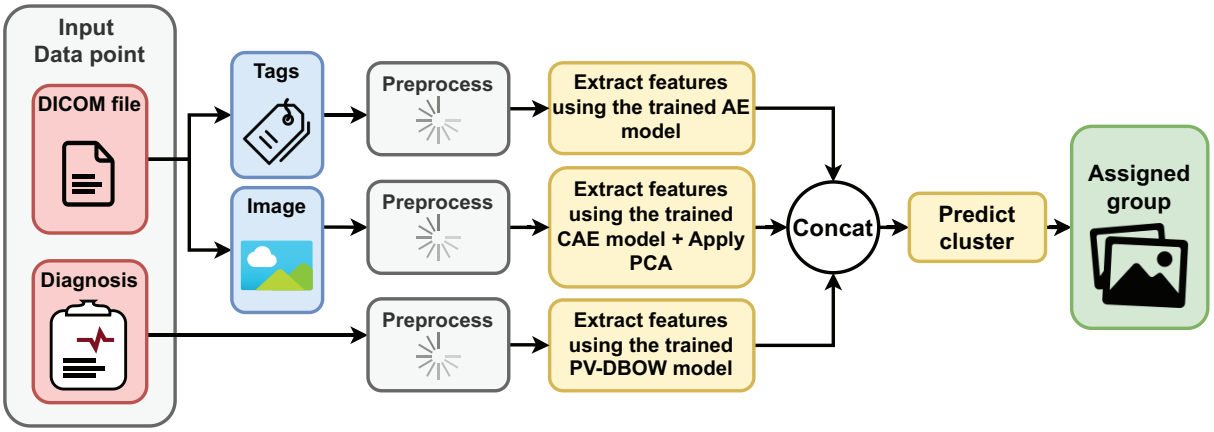



**Figure 2.13:** The fully unsupervised labelling algorithm. From an example data point, each data source was processed independently and then fused together to form a single embedding. Afterwards, this embedding was used to assign this data point to a group [41].

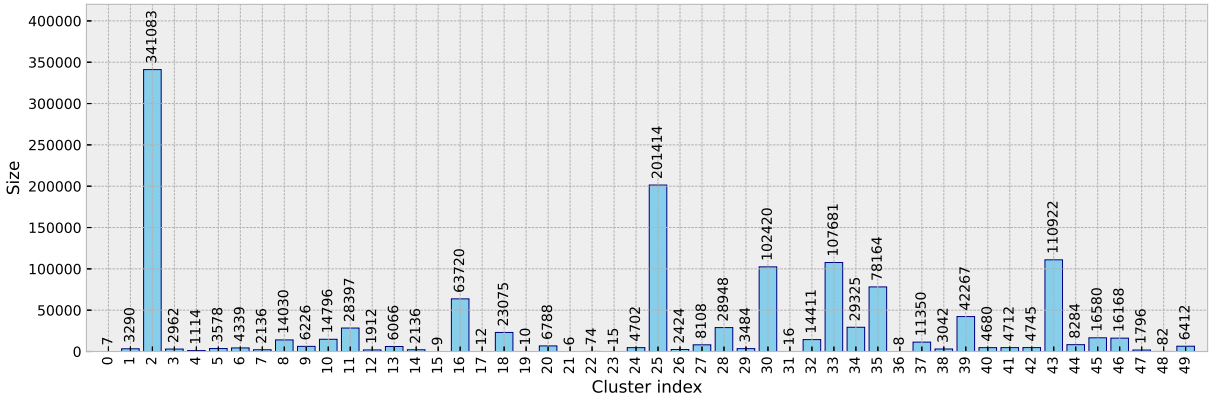



**Figure 2.14:** Sizes of obtained groups in the pseudo-labelled RadiologyNET dataset [41].

ical areas are frequently grouped together. This phenomenon is particularly noticeable in the torso region, where discerning between the abdomen, gastrointestinal tract, and pelvis can be challenging (e.g. cluster 33), as well as between the abdomen and chest (e.g. cluster 2). Such groupings can be attributed to the difficulty in distinguishing precise anatomical boundaries within body regions, as it is not unusual that a single study contains multiple parts of the torso – for example, an MR image capturing both the abdomen and the pelvis. A similar pattern is observed in spine groupings, where different spinal segments are often clustered together (e.g. clusters 30 and 37). Additionally, images of body extremities (e.g. hands and feet) were frequently grouped together (cluster 49), despite their visual dissimilarity and differences in anatomical location. On the other hand, clusters 16, 32, 35 and 39 displayed evidence of containing outliers, as they included images depicting anatomically unrelated regions (e.g. leg, abdomen, head, urinary tract).

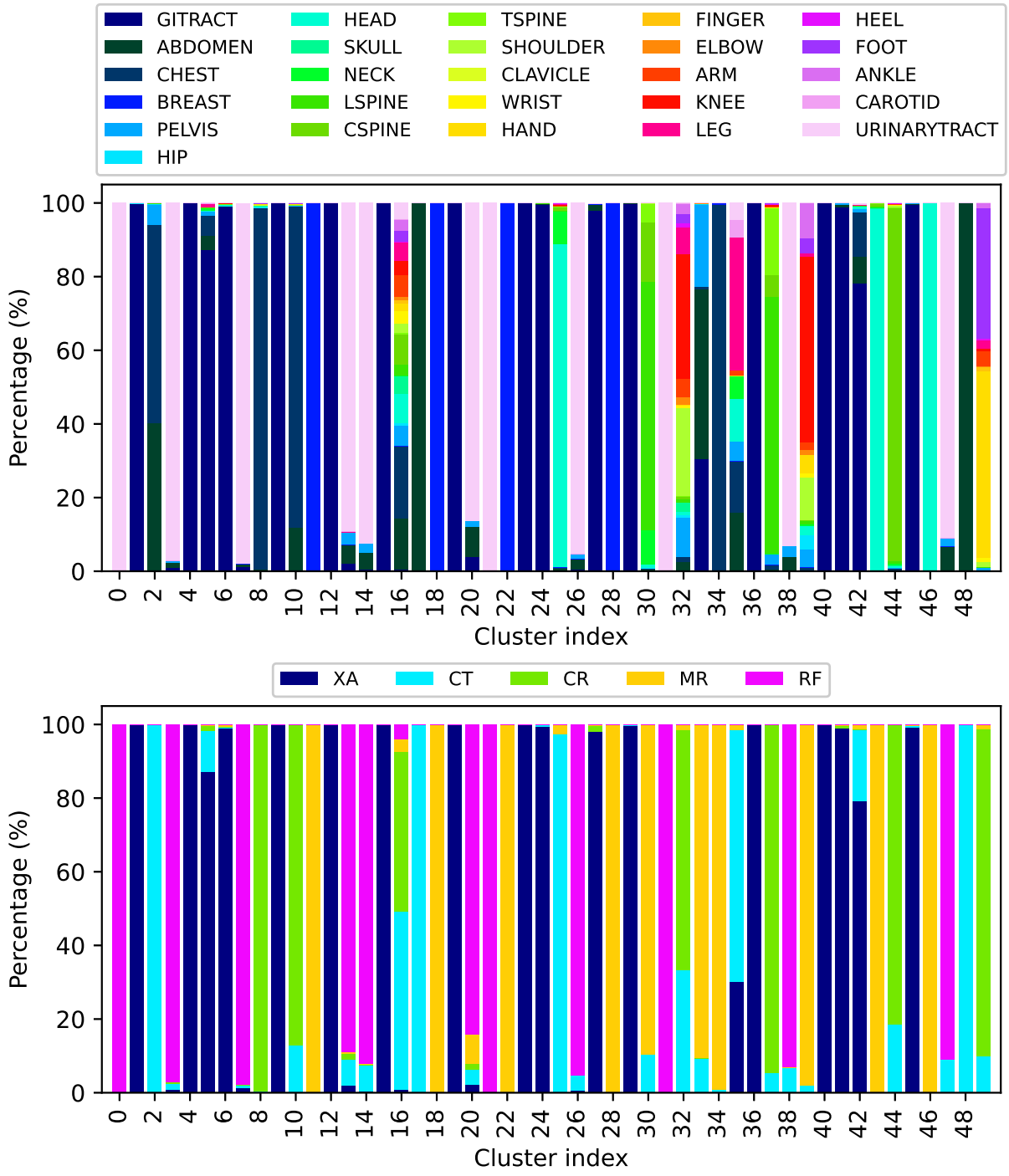



**Figure 2.15:** Grouping quality regarding modality and examined body part, for the labelled RadiologyNET dataset. The first image shows how homogeneous the clusters are when observing the body part, while the second one shows the different modalities in each cluster [41].

Such heterogeneity suggests that these clusters may contain anomalies or cases where the anatomical or diagnostic context is ambiguous.

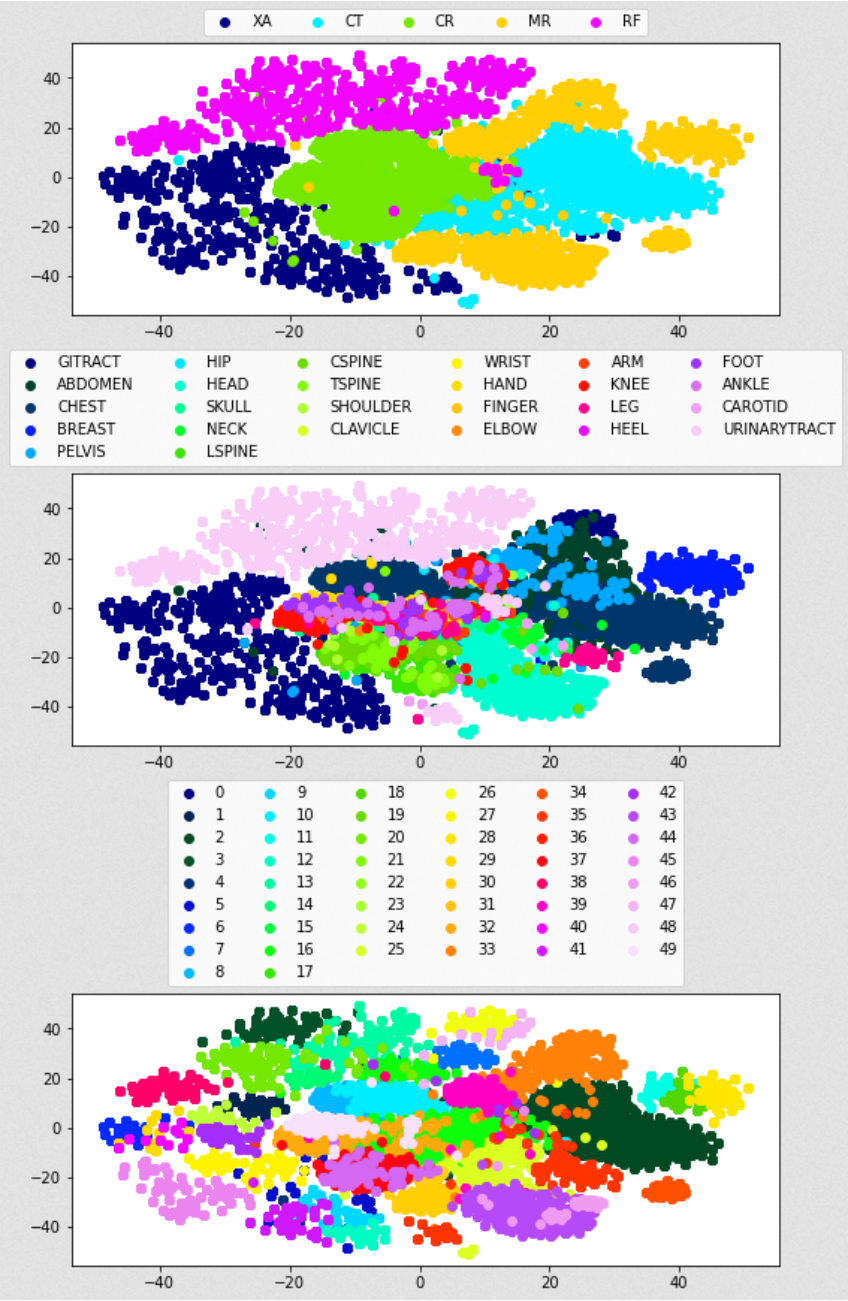



**Figure 2.16:** t-SNE visualisation of the generated embeddings, showing clustering patterns by imaging modality (top), examined body part (centre), and generated pseudo-labels (bottom).

The discussed clustering patterns and tendencies can also be seen in the t-SNE visualisation shown in Figure 2.16, which depicts the embeddings with respect to modality,

examined body part, and the generated pseudo-labels.

After the pseudo-labels were generated, they were used for supervised pretraining of RadiologyNET foundation models which is described in the following chapter.

# 3. Chapter

# RADIOLOGYNET PRETRAINING AND EXPERIMENTAL DESIGN

Transfer learning, in the context of deep learning, leverages knowledge from models pretrained on large datasets to improve training progress and stability on downstream tasks. This process is particularly valuable in medical imaging, where obtaining large labelled datasets can be challenging, and therefore, transferring knowledge from previously trained models can serve as a good starting point for downstream tasks (that are usually resource-limited).

The process of TL is shown in Figure 3.1. In the context of TL, pretraining refers to the initial phase of model training where the network learns a set of feature representations from a large (ideally diverse) dataset. After pretraining the models, the next step is to evaluate their effectiveness in TL by testing their performance on various downstream tasks. In medical TL, these downstream tasks vary in complexity and may involve classification, regression, or segmentation, depending on the nature of the medical images and the clinical questions being asked.

- **Classification**. The objective is to assign a specific label to an input image based on its visual features. In medical imaging, classification tasks often involve identifying the presence or absence of a specific pathology.
  - **Binary Classification**. The model predicts one of two possible classes, such as “disease present” or “disease absent”. This type of classification is particu-

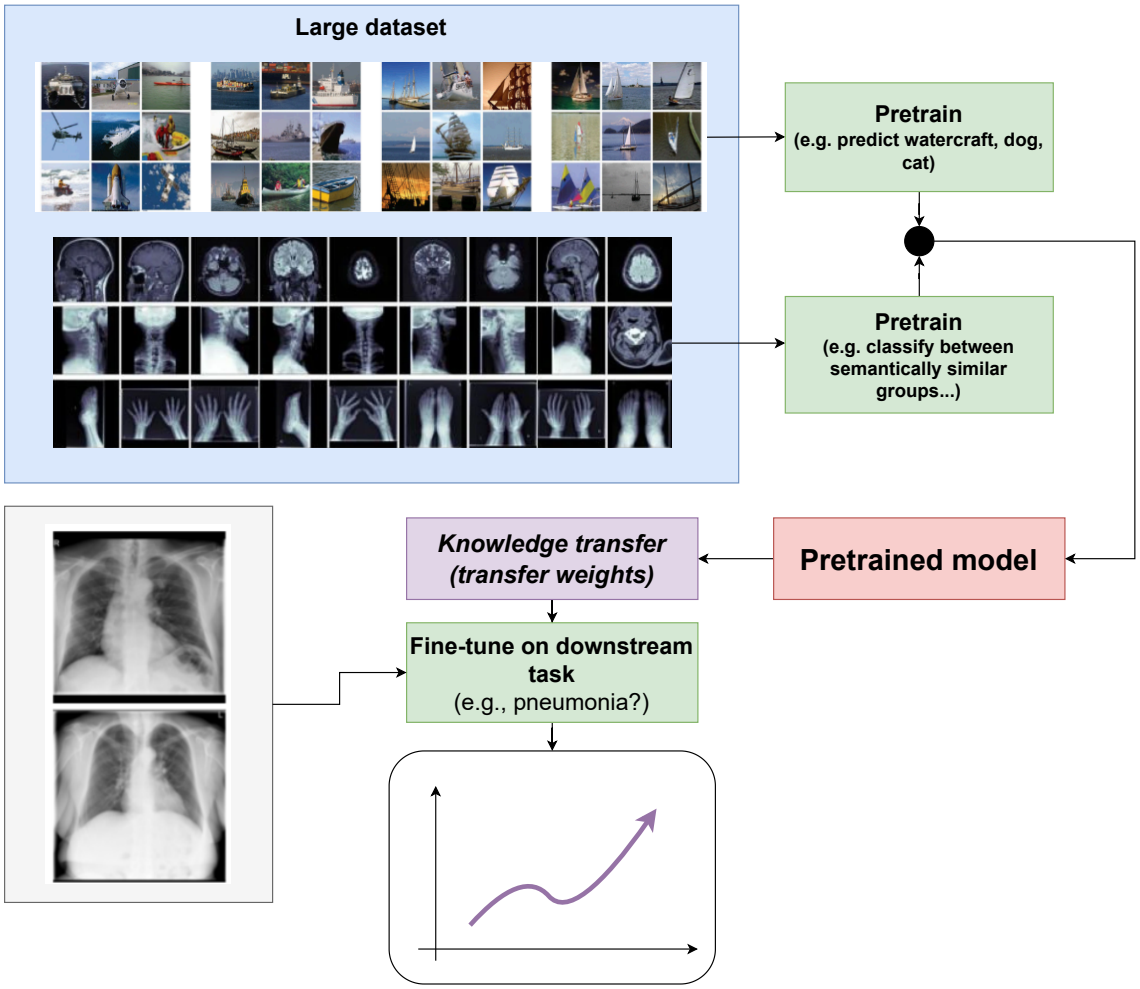



**Figure 3.1:** The general process of pretraining, transfer learning and fine-tuning on downstream tasks.

larly useful for diagnostic tasks where the goal is to determine the presence of a specific condition (e.g. detecting osteopenia in wrist X-ray images).

- **Multiclass Classification**. The model assigns an image to one of several possible classes. This approach is useful when differentiating between multiple conditions or subtypes of a disease (e.g. classifying different types of brain tumours on MR images).

- **Regression**. In regression tasks, the model predicts a continuous numerical value rather than a discrete label. An example in medical imaging can include estimating the skeletal age based on an X-ray image.

– **Segmentation**. The goal is to partition an image into meaningful regions, usually by labelling each pixel (or voxel, if working with three-dimensional volumes) according to the anatomical structure or pathological area it represents. In radiological applications, segmentation is commonly used to distinguish between organs, lesions, or tumours (e.g. segmenting the liver from an abdominal MR image, or identifying cancerous cells in a CT image).

The first part of this chapter describes the process of pretraining foundation models on the RadiologyNET dataset. After describing the process of pretraining, the process of TL and fine-tuning is detailed, as is the performed statistical analysis.

## 3.1. Model Pretraining

This section details the process of pretraining RadiologyNET foundation models and is structured as follows. Firstly, the dataset preparation process for pretraining is described, including image export and data filtering. The next subsections outlines the pretraining setup, covering the selection of model architectures, hyperparameter optimisation, and the computational environment (i.e. the available hardware). In this section, it is also described that models were pretrained on two distinct task types: (i) classification and (ii) reconstruction. Additionally, pretraining was conducted on different subsets of the RadiologyNET dataset: (i) the primary models were pretrained on the entire multi-modality dataset, while (ii) several models were also pretrained on single-modality data (e.g. CT-only pretrained models). Finally, the use of ImageNet-pretrained models is detailed.

### 3.1.1. Preparing the dataset for pretraining

The dataset used in this phase is the one annotated and grouped as described in the previous chapter, where an unsupervised labelling algorithm was built and used to generate pseudo-labels. These pseudo-labels, presented in Figures 2.14 and 2.15, represent distinct groups of semantically similar data points, forming the basis for supervised pretraining tasks. Some of the clusters obtained in the automated annotation process exhibited high heterogeneity, and any groups where the entropy of obtained labels exceeded an empirically determined threshold were excluded from further use. As a result,

14 clusters were removed from the dataset, leaving a total of 36 pseudo-labels suitable for the pretraining task.

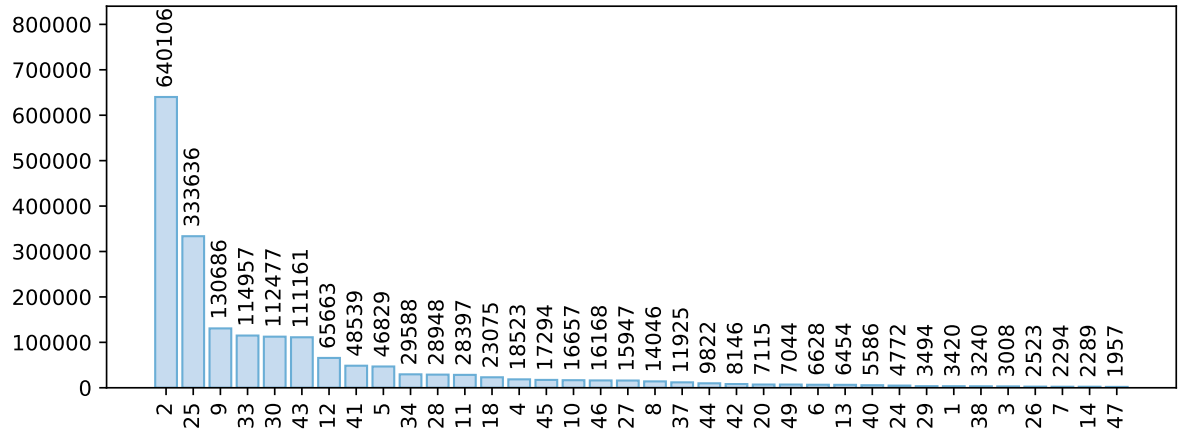



**Figure 3.2:** Cluster (pseudo-class) distribution in the dataset used for pretraining.

To prepare data for pretraining of RadiologyNET foundation models, images were exported to an 8-bit format, to match ImageNet's colour depth. The windowing process used for this 8-bit conversion follows the same methodology described in Section *2.3.2.* It is important to note that the data points shown in Figure 2.14 represent individual DICOM files. However, a single data point can contain multiple images, as a single DICOM file may hold several slices or frames (e.g. CT images can be three-dimensional volumes). Consequently, multiple PNG images can be extracted from each DICOM file. Processing all valid data points led to a total of 1,902,414 PNG images, with sizes of each group shown in Figure 3.2.

The distribution of medical imaging modalities present in the pretraining dataset is shown in Figure 3.3a. The dataset encompasses a wide range of anatomical regions and body parts, including hands, ankles, the abdomen, and the brain. Figure 3.3b presents the distribution of the *BodyPartExamined* attribute, as recorded in the DICOM file headers. Although this attribute is manually entered by physicians and is therefore susceptible to errors, it can still offer insight into the anatomical diversity of the dataset. As depicted in Figure 3.3, the pretraining dataset consisted mostly of chest, abdominal and head images, captured mostly using MR and CT.

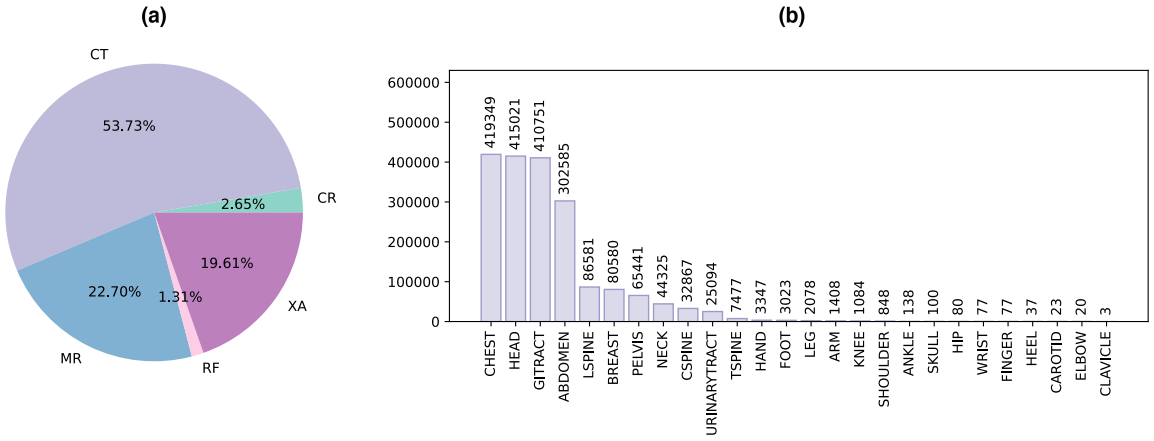



**Figure 3.3:** The overall distribution of different imaging modalities (a) and anatomical regions (b) found in the dataset used for pretraining.

### 3.1.2. Pretraining setup

A variety of neural network architectures were employed for pretraining, chosen based on their previous proven effectiveness in various medical imaging tasks [4, 5, 55, 93]. The selected architectures included CNNs and segmentation networks. A comprehensive list of pretrained models used in the experiments is presented in Table 3.1.

The primary objective of the pretraining phase was to train models to predict one of the 36 classes shown in Figure 3.2. These classes correspond to the pseudo-labels obtained through unsupervised annotation and represent semantically similar groups of medical imaging data. An exception to this classification task was the U-Net architecture which was used for segmentation tasks, and was therefore pretrained as a reconstruction-based task. In this type of pretraining, the aim was to reconstruct the initial (input) image, with the aim that features learned during reconstruction (such as textures and patterns) would help in downstream segmentation tasks. To summarise, all models provided in Table 3.1 were pretrained as **classification** tasks using the generated pseudo-labels, except for U-Net, which was pretrained as a **reconstruction** task.

The pretraining process incorporated data augmentation to enhance model generalisation. Augmentation techniques included rotation, scaling, flipping, and intensity variation. Due to different group sizes within the dataset (which is visible in Figure 3.2), oversampling techniques were applied to mitigate the potential impact of class imbalance. Learning rates were selected based on the optimal values reported in the original studies

**Table 3.1:** Overview of neural network architectures pretrained in this research.

| Architecture | Pretrained on | Overview |
|---|---|---|
| ResNet18 | Multi-modality, CR-only | Known for robust feature extraction using residual connections. Widely used in image processing [30]. |
| ResNet34 | Multi-modality, CR-only | |
| ResNet50 | Multi-modality, CT-only, MR-only. | |
| EfficientNetB3 | Multi-modality, CR-only | Scales depth, width, and resolution efficiently. Achieves high accuracy with few parameters [32]. |
| EfficientNetB4 | Multi-modality | |
| VGG16 | Multi-modality | Classic deep convolutional network with uniform layers. Widely used for image classification, feature extraction [51], etc. |
| InceptionV3 | Multi-modality, CR-only | Incorporates multi-scale feature extraction using inception modules [50]. |
| DenseNet121 | Multi-modality, CR-only | Uses dense connectivity for improved gradient flow [31]. |
| MobileNetV3Small | Multi-modality, MR-only | Optimised for mobile and embedded vision applications. Lightweight and efficient thanks to depth-wise separable convolution [52]. |
| MobileNetV3Large | Multi-modality, CR-only | |
| U-Net | Multi-modality, ImageNet | Designed for medical image segmentation. Uses an encoder-decoder architecture with skip connections [73]. |

of each architecture [30, 31, 32, 50, 51, 52, 73]. Early stopping was implemented to stop training when the validation loss did not improve for 10 epochs, with validation being performed after every 5% of training data was processed (meaning that validation was performed 20 times per epoch).

The pretraining process was conducted on a server equipped with four NVIDIA RTX A6000 GPUs, each with 48 GB of VRAM, and a system with 512 GB of CPU RAM. Due to the substantial size of the RadiologyNET dataset and the complexity of the models, pretraining typically required approximately seven days per model. Multiple models were often trained concurrently (in parallel).

### 3.1.3. Single-modality pretraining

Primary models were pretrained on the entire multi-modality dataset to exploit the diversity of imaging modalities and anatomical regions. However, to evaluate whether TL benefits more from diverse data or from modality-specific data, additional models were pretrained on single-modality subsets. The goal of this experiment was to determine whether using modality-aligned pretrained models would obtain better performance compared to multi-modality pretrained models, thus testing the importance of pretraining data diversity. For example, when evaluating downstream tasks involving MR images, models were pretrained exclusively on MR data, and data from other modalities were excluded (in this example, only images captured in the same modality as those shown in Figure 2.5c were used).

The choice of modality for each architecture was guided by the specific downstream tasks, which will be described in the following chapter. Consequently, not all architectures were pretrained for each single modality; and the architectures that were pretrained on single-modality data are specified in Table 3.1 and emphasised in different colours. For single-modality pretraining, only images from the selected modality were used, while data from other modalities were excluded.

### 3.1.4. ImageNet pretraining

As is visible in Table 3.1, U-Net was (in addition to being pretrained on the RadiologyNET dataset) also pretrained on ImageNet, due to the lack of publicly available

ImageNet-pretrained weights for U-Net. For this reason, the entire ImageNet dataset was downloaded and a pretraining pipeline built for U-Net ImageNet pretraining. This ImageNet-pretrained U-Net was pretrained as a reconstruction task.

For the remaining architectures shown in Table 3.1, ImageNet-pretrained weights were directly obtained from the official PyTorch repositories. These weights were selected as they represent the standard initialisation used in previous studies.

## 3.2. Transfer Learning and Fine-Tuning

After selecting the downstream tasks, three approaches were tested to evaluate model performance: (i) training from randomly initialised weights, (ii) fine-tuning on ImageNet, and (iii) fine-tuning on RadiologyNET. For simplicity, any model trained from scratch (i.e. with randomly initialised weights) will henceforth be referred to as *Baseline*. The experimental setup was designed to closely follow the configurations proposed in previous studies for similar tasks, including model architecture, optimiser, and loss functions. Additionally, task-specific adjustments (e.g. incorporating auxiliary information when/where relevant) were implemented to ensure consistent evaluation across approaches.

It is important to note that the primary aim of this research was not to achieve state-of-the-art performance, but to systematically compare the effectiveness of different TL strategies and pretrained models, while making use of network architectures that performed well on the tasks in previous studies.

All models were trained in equal conditions, with the training process running for a maximum of 200 epochs, with an early stopping criterion applied if the validation performance did not improve over 10 consecutive epochs. During training, all model parameters were unfrozen and subject to fine-tuning. Although partial (un)freezing of model parameters was considered and tested, the results did not yield any improvements. Regarding the train-valid-test splits, some publicly available datasets already have a dedicated train, validation and test subsets, and if those were available, then those existing subsets were used. Otherwise, if pre-determined subset splits were unavailable, the train, validation and test subsets were randomly split in a $75\% : 12.5\% : 12.5\%$ ratio.

To account for variability in model optimisation, each model was trained five times using learning rates logarithmically sampled from the range $[10^{-2}, 10^{-5}]$, with a base-

10 step size. Initially, higher ($\geq 0.1$) and lower ($\leq 10^{-6}$) learning rates were also tested. However, they were later excluded due to consistently lower performance across all models and challenges. Models trained with lower learning rates often failed to converge within the specified number of epochs, while those with higher learning rates demonstrated worse overall performance.

After training multiple models for each downstream challenge, hyperparameter selection was guided by their performance on the validation set. Specifically, the optimal hyperparameter configuration for each model architecture was identified as the one that achieved the highest evaluation metric on the validation subset. Once the best-performing configuration was selected, the corresponding model using those hyperparameters was subsequently evaluated on the test set. Final statistical analyses were conducted using these test set results, meaning that all models were evaluated under their most effective settings on unseen data.

Downstream tasks were evaluated on the same machine used to build the unsupervised pipeline, which was previously described in Section 2.4.2.. However, this time the server was upgraded with 2× NVIDIA L40S (which was shortly thereafter expanded to 4× L40S GPUs), each with 48 gigabytes of VRAM memory. Completing all processes required for a single challenge took several days, which includes TL from three approaches, on multiple learning rates and neural network topologies, in five independent runs. The total time varied between architectures, with smaller networks requiring significantly less time than larger neural networks.

## 3.3. Metrics and Statistical Analysis

In classification machine learning tasks, the predictions can be sorted into the following categories: true positive (TP), true negative (TN), false positive (FP) and false negative (FN). True positives and true negatives together represent the total number of correct predictions, while false positives and false negatives add up to the total number of incorrect predictions. The overall accuracy of a classification tasks is therefore calculated as:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.1}$$

In addition to the overall accuracy of a model, it is possible to calculate its precision, which can be considered equivalent to answering *of all instances predicted as positive, how many were actually positive?*; and recall, i.e. *of all actually positive examples, how many were predicted as positive?*. Precision and recall are formulated as:

$$precision = \frac{TP}{TP + FP}, \qquad recall = \frac{TP}{TP + FN}. \tag{3.2}$$

Precision and recall are frequently used metrics in machine learning classification tasks. They are also often combined into a single metric: the F1-score. The F1-score is calculated as the harmonic mean of precision and recall:

$$\text{F1-score} = \frac{2 \cdot precision \cdot recall}{precision + recall} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}. \tag{3.3}$$

Similarly, semantic segmentation tasks can be considered as *pixel-wise* classification, i.e. each pixel is attributed to a specific class. Commonly used metrics in segmentation tasks include Intersection-over-Union (IoU) and Dice-Sørensen coefficient [94], with the latter often abbreviated as *Dice* score. They are formulated as follows.

$$Dice = \frac{2 \cdot \text{Area of Overlap}}{\text{Total Area}} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}, \tag{3.4}$$

$$IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}} = \frac{TP}{TP + FP + FN}. \tag{3.5}$$

The Dice score is mathematically equivalent to the F1-score given in Equation 3.3, but applied on a per-pixel basis in semantic segmentation tasks. All of the previously presented metrics (accuracy, precision, recall, F1-score, IoU and Dice) have a maximal score of 1 (indicating perfect overlap with the ground truth), with 0 being the minimal possible obtainable value (completely incorrect predictions).

Regression-based tasks use different metrics to measure the goodness-of-fit. Commonly used metrics include Mean Absolute Error (MAE), MSE and Root Mean Squared Error (RMSE). They are calculated by measuring the distance between the predicted value $\hat{y}$ and the ground truth $y$:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|, \tag{3.6}$$

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{3.7}$$

where $n$ is the number of instances. RMSE is the square root of MSE and penalises larger errors more than MAE.

For the segmentation task, IoU and the Dice score were calculated to measure the quality of the predicted masks [94]. In the regression task, MAE and RMSE were used to quantify the difference between predicted and actual values. Classification tasks were evaluated using Accuracy and F1-score. Out of the listed metrics, the Dice score was used for statistical tests on segmentation tasks, MAE was the primary metric for comparison on regression tasks, and the F1-score was used for statistical comparison of the classification-based tasks.

To evaluate the statistical significance of the obtained results, a Levene test [95] was first performed to determine whether significant differences in variance exist. If the test indicated significant variance differences ($p < 0.05$), non-parametric Kruskal-Wallis tests [96] were used to compare the performance across three or more groups. Where the Kruskal-Wallis test showed potential significant differences, pairwise comparisons were computed using the Mann-Whitney U (MWU) test, with Bonferroni correction applied to account for multiple comparisons. If the Levene test did not indicate significant variance differences, parametric one-way Analysis of Variance (ANOVA) would be performed when comparing three or more groups. If the ANOVA test identified significant differences, post-hoc pairwise comparisons would be conducted using the Tukey's Honestly Significant Difference test. If the Levene's test did not indicate significant difference in variance between two groups, then an Independent instances t-test would be performed.

The initial results revealed no significant performance differences between models pre-trained on ImageNet and those pretrained on RadiologyNET. To further investigate potential differences in resource efficiency, the `codecarbon` package [97] was implemented to measure emissions and energy consumption during training for selected downstream tasks. As expected, the recorded emissions were strongly correlated with the number of

training epochs required to reach convergence. Given this direct relationship, the number of epochs was adopted as an additional evaluation metric and is reported in the results. More details on the chosen downstream tasks and the obtained results is given in the following chapter.

# 4. Chapter

# THE EFFICACY OF RADIOLOGY-NET FOUNDATION MODELS

Following the pretraining phase, RadiologyNET foundation models were evaluated across a variety of downstream tasks to evaluate their effectiveness in TL. These datasets were chosen to ensure diversity in radiological imaging modalities, anatomical regions, and task types. The selected tasks encompass segmentation, regression, binary classification, and multiclass classification, covering a wide range of medical imaging problems and challenges.

This chapter is structured as follows. First, each of the selected downstream tasks is described, along with the motivation for their inclusion. The results of TL evaluation using RadiologyNET models are then presented, with performance compared against ImageNet-pretrained and randomly initialised (*Baseline*) models. Given the minimal performance differences observed, the evaluation was extended to resource-limited conditions by limiting both training time and training data. Finally, the implications of these findings are discussed. The complete experimental workflow, from the pretraining phase to transfer learning, fine-tuning, and evaluation, is depicted in Figure 4.1.

## 4.1. Chosen Challenges

The selected challenges were: LUng Nodule Analysis Challenge (LUNA) – *segmentation*; Radiological Society of North America (RSNA) Pediatric Bone Age Challenge (PBA)

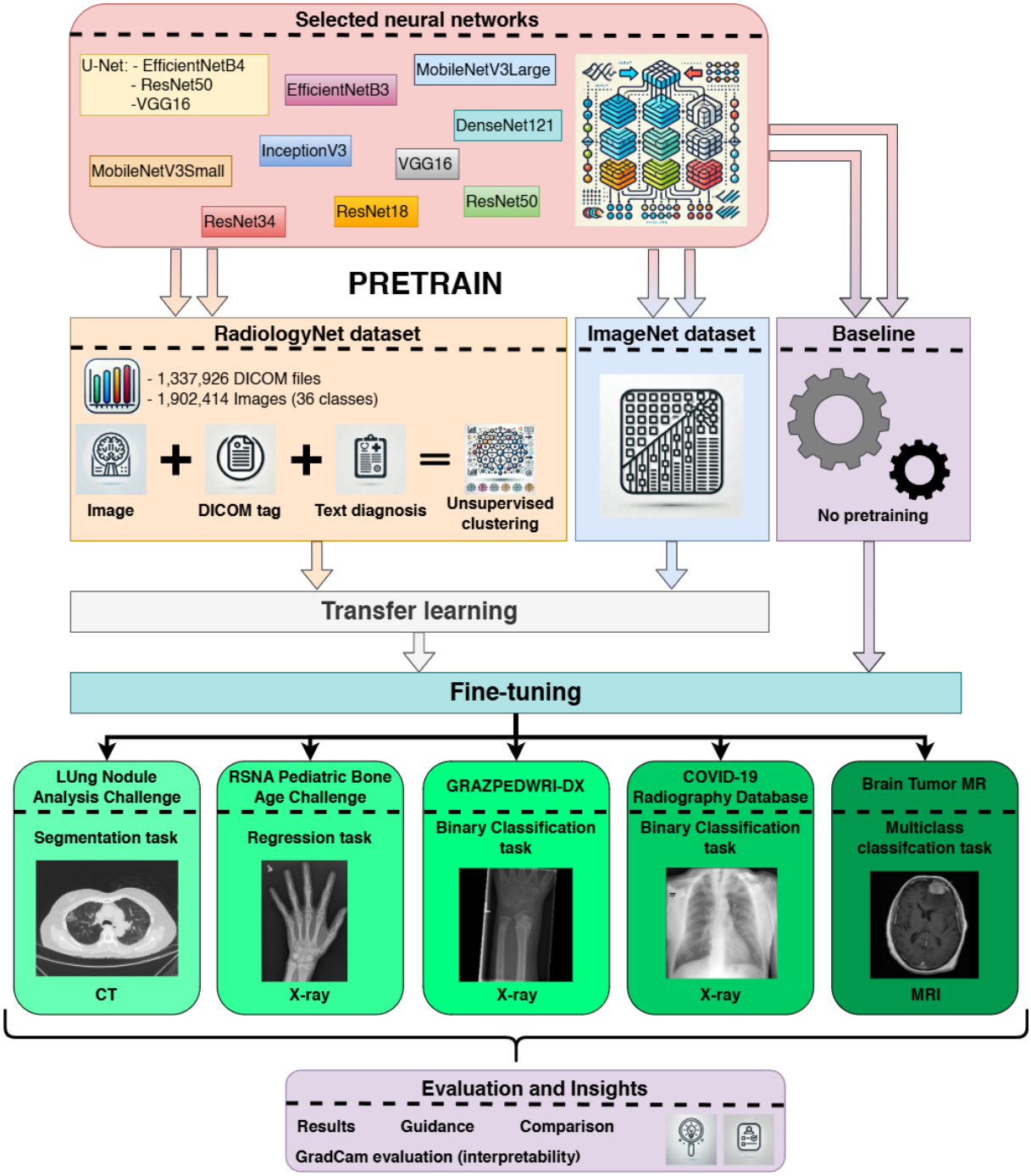



**Figure 4.1:** A workflow of the conducted experiment, from the pretraining phase to transfer-learning, fine-tuning and evaluation on the downstream tasks.

– *regression*; GRAZPEDWRI-DX and COVID-19 – *binary classification*; and Brain Tumor MRI (BTMR) – *multiclass classification*). The selected challenges covered different problem types, and as the RadiologyNET dataset is imbalanced with regard to imaging modalities and anatomical regions (Figure 3.3), the datasets were chosen to include (i) data which aligns with the pretraining dataset's domain, and (ii) data which shows less overlap with the original pretraining dataset. It is important to note that *overlap* in this context refers only to domain relevance, and that the RadiologyNET dataset and downstream tasks were completely independent, i.e. the patient scans found in downstream tasks were not a part of the RadiologyNET dataset. All of the downstream models were trained using the process and hyperparameters previously described in Section *3.2.*

Example images from each dataset can be seen in Figure 4.1. The list of metrics, neural network topologies and pretraining domains used for each challenge is given in Table 4.1. Metrics used for statistical tests are emphasised. The following subsections detail the selected publicly available datasets and the architectures evaluated for each dataset.

## LUng Nodule Analysis Challenge (LUNA)

The LUng Nodule Analysis Challenge (LUNA) [57] is based on the Lung Image Database Consortium and Image Database Resource Initiative (LIDC-IDRI) dataset [56], which contains a total of 1,018 CT lung scans. The challenge comprises two tasks: (i) nodule classification, where the objective is to determine whether specified locations in a scan correspond to nodules, and (ii) nodule segmentation, where the goal is to generate a mask from a full CT scan that marks the nodule regions. For this study, the segmentation task was selected.

The winning solution for the LUNA segmentation task relied on the U-Net architecture [73]. Motivated by the success of the U-Net-ResNet50 variant [98], which replaces the encoder branch with ResNet50, this research used several other classification-pretrained models in a U-Net-like topology to perform nodule segmentation. The selected architectures included VGG16 [51], EfficientNetB4 [32, 99], and the aforementioned ResNet50 [30]. To preserve U-Net's signature multi-resolution feature fusion, skip connections were used to link outputs from intermediate encoder layers to their corresponding decoder

**Table 4.1:** Overview of downstream tasks, task types, tested architectures, pretraining data (single- or multi-modality), and evaluation metrics. Emphasised metrics were used for statistical tests and comparisons.

| Dataset | Task Type | Tested Architecture | Pretraining data | Evaluation Metric(s) |
|---|---|---|---|---|
| RSNA PBA | Regression | EfficientNetB3 | Multi-modality, CR-only | **MAE**, RMSE, **Epochs** |
| | | InceptionV3 | Multi-modality, CR-only | |
| COVID-19 | Binary Classification | MobileNetV3Large | Multi-modality, CR-only | **F1-score**, Accuracy, **Epochs** |
| | | ResNet18 | Multi-modality, CR-only | |
| GRAZPEDWRI-DX | | DenseNet121 | Multi-modality, CR-only | |
| | | ResNet34 | Multi-modality, CR-only | |
| BTMR | Multiclass Classification | MobileNetV3Small | Multi-modality, MR-only | |
| | | ResNet50 | Multi-modality, MR-only | |
| LUNA | Segmentation | U-Net | Multi-modality, ImageNet | **Dice**, IoU |
| | | U-Net-VGG16 | Multi-modality | |
| | | U-Net-EfficientNetB4 | Multi-modality | |
| | | U-Net-ResNet50 | Multi-modality, CT-only | |

layers.

The LUNA dataset consists of CT images of the lungs. When comparing this dataset with the distribution of modalities and anatomical regions present in the RadiologyNET pretraining dataset (Figure 3.3), it can be observed that there is a significant domain overlap.

## Pediatric Bone Age Challenge (PBA)

The RSNA Pediatric Bone Age dataset [55] comprises 14,236 hand radiographs labelled by expert radiologists. The primary objective is to estimate skeletal age (a regression task) where the predicted output represents bone age expressed in months.

The winning solution in this challenge used the InceptionV3 architecture [50], where the network output was concatenated with the sex information also provided in the public dataset. This concatenated feature vector was then passed through additional dense layers to produce the final bone age prediction. Motivated by the success of CNNs in this challenge [55], this study used EfficientNetB3 following a similar approach. Specifically, the output from EfficientNetB3 was concatenated with the available sex information before being processed through fully connected layers to predict skeletal age.

It is important to note that this dataset consists exclusively of CR images of hands, while only 2.65% (Figure 3.3) of the RadiologyNET pretraining dataset is comprised of CR images (and images of hands are also scarce compared to other body regions). Therefore, this dataset exhibits limited domain alignment with the pretraining dataset. This could be a potential challenge for effective TL, as patterns learned during pretraining might not resemble those found in this downstream task.

## GRAZPEDWRI-DX

The GRAZPEDWRI-DX dataset [53] contains 20,327 digital radiographs of wrists, annotated by expert radiologists. These annotations are suitable for various detection and classification tasks; however, for the purposes of this study, the task of osteopenia classification was selected, defining this challenge as a binary classification problem. Among the available images, 2,473 are labelled as showing osteopenia. To address the class imbalance, undersampling of non-osteopenia cases was performed, resulting in a balanced

dataset of 4,946 images.

Previous studies [93] have demonstrated the effectiveness of various ResNet and DenseNet architectures in classifying osteopenia using the GRAZPEDWRI-DX dataset. These findings guided the selection of similar CNN-based architectures for the current study.

It is important to note that, similarly to the RSNA Pediatric Bone Age dataset, GRAZPEDWRI-DX contains only wrist radiographs (CR images). Since wrist radiographs are scarce in the RadiologyNET pretraining dataset (Figure 3.3), this results in limited domain overlap between the pretraining and downstream datasets, which also marks this dataset as challenging for RadiologyNET transfer learning.

## COVID-19 Radiography Database

The COVID-19 Radiography database [4, 5] consists of chest CR images from patients diagnosed with COVID-19 (3,616 images) alongside normal chest radiographs (10,192 images). While the dataset also includes images depicting lung opacity (non-COVID-19 lung infection) and viral pneumonia cases, the current study focuses on the binary classification task of distinguishing between COVID-19 and normal cases. Due to the inherent class imbalance between normal and COVID-19 cases, similarly to process performed in the GRAZPEDWRI-DX dataset, the normal cases were undersampled to match the number of COVID-19 cases.

The original research [5] evaluated several popular architectures, identifying ResNet18 as one of the best performers, with MobileNetV2 achieving a comparable score (0.01% difference). As a part of this research, the newer MobileNetV3Large [52] was selected alongside ResNet18 [30] for evaluation.

Although this dataset primarily comprises radiographs, which are relatively scarce in the RadiologyNET pretraining dataset, it consists of chest images. Since chest imaging represents the most prevalent anatomical regions within the RadiologyNET dataset (as it can be seen in Figure 3.3b), this downstream dataset serves as a midpoint in terms of domain alignment with the pretraining data.

### Brain Tumor MRI (BTMR)

The Brain Tumor MR Imaging dataset [54] consists of 7,023 MR images of the brain where the objective is to correctly identify the tumour type (or its the absence). The dataset is annotated with four class labels: glioma, meningioma, pituitary, and no tumour, making this a multiclass classification problem. This dataset originated from a Kaggle competition, which featured submissions in various popular network topologies, including MobileNet architectures [52] and ResNet variants [30].

Among the imaging modalities present in the RadiologyNET pretraining dataset, MR is the second most prevalent (accounting for 22.7% of the data), and images of the head were among the most common in the pretraining dataset (second only to chest images, as shown in Figure 3.3b). Therefore, this downstream dataset exhibits a high degree of domain alignment with the pretraining data. This strong alignment suggests that the patterns learned during pretraining are likely to contribute positively during fine-tuning on this dataset.

## 4.2. Results

The results presented in this section were obtained on the test subsets of each downstream task, using models that demonstrated the best performance on the validation subset of each respective dataset. Validation results are provided in the Appendix, specifically in Tables A.1, A.2, A.3, and A.4. The only exception is LUNA, where the reported results (shown in Table 4.2) were computed on the validation subset (as reported in [100]).

The best-performing models for the PBA, GRAZPEDWRI-DX, COVID-19, and BTMR datasets are summarised in Tables 4.3, 4.4, 4.5, and 4.6, respectively. Each table presents the average performance across five independent runs, along with the best recorded performance on the test subset. The exact $p$-values for metric comparisons are reported in the Appendix, in Table A.5, while the comparison of training length (in terms of the number of epochs) is given in Table A.6.

Additionally, examples of model predictions for the BTMR, GRAZPEDWRI-DX, COVID-19, and PBA datasets are presented in Figure 4.2. The figure shows both consensus cases (where all models correctly predicted the target class) and more challenging

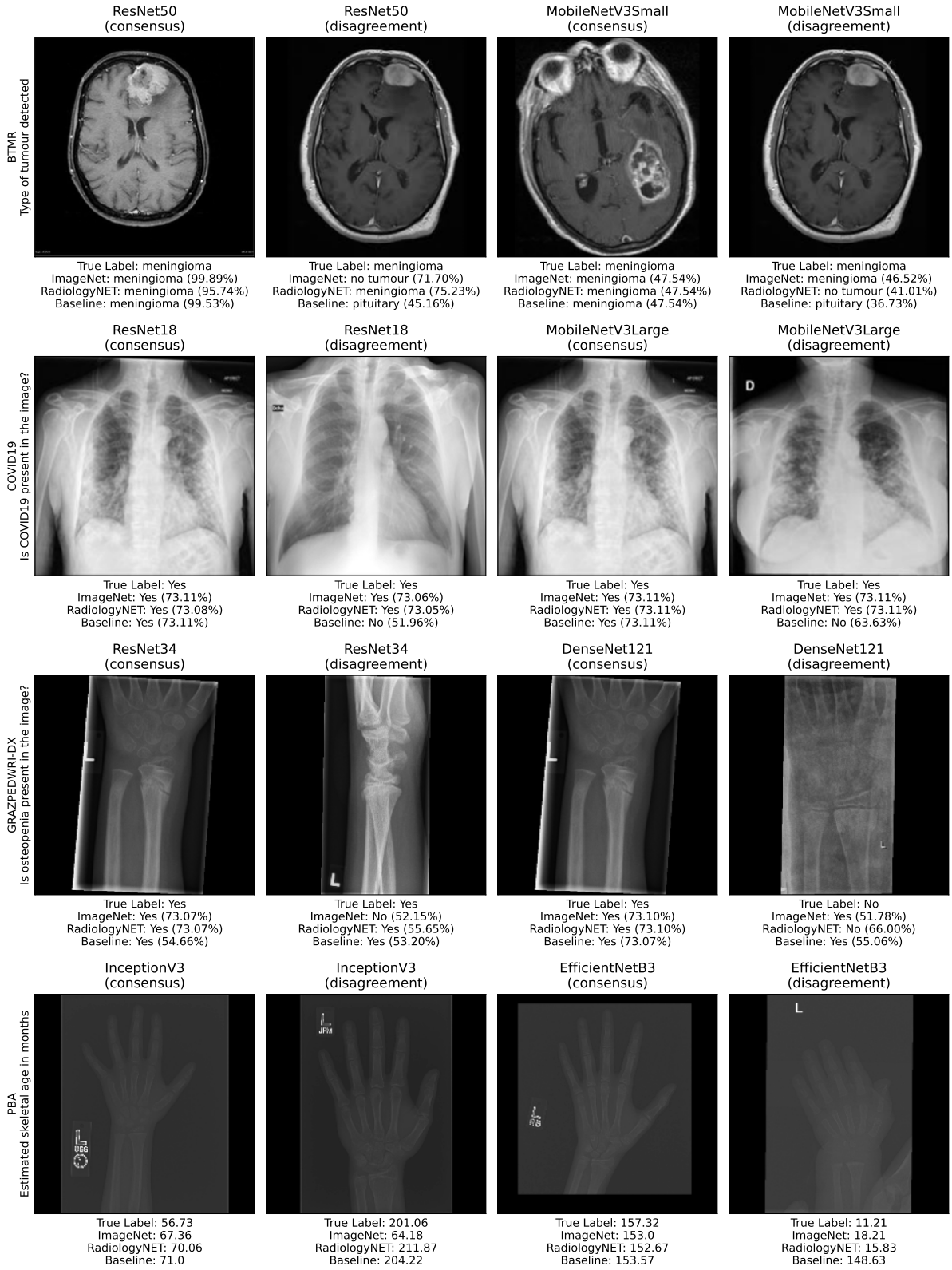



**Figure 4.2:** Mosaic of example predictions illustrating (dis)agreement among models. For classification tasks, the predicted class is accompanied by its associated probability (i.e. softmax output), which implies the models' confidence in their predictions.

instances where predictions were inconsistent or incorrect (the models had different predictions).

## LUng Nodule Analysis Results

**Table 4.2:** Results on the LUNA dataset are shown for U-Net, U-Net-ResNet50, U-Net-EfficientNetB4, and U-Net-VGG16 models; for Reconstruction (R) and Classification (C) pretraining strategies. Best results are emphasised.

| | **Pretrain Strategy** | **TL Model** | **LR** | **Dice Score** | **IoU** |
|---|---|---|---|---|---|
| U-Net | R | ImageNet | $10^{-4}$ | 0.111 ± 0.002 | 0.063 ± 0.002 |
| | R | RadiologyNET | $10^{-4}$ | 0.111 ± 0.002 | 0.063 ± 0.001 |
| | N/A | Baseline | $10^{-4}$ | **0.616 ± 0.012** | **0.500 ± 0.011** |
| U-Net-EfficientNetB4 | C | ImageNet | $10^{-4}$ | 0.685 ± 0.016 | 0.582 ± 0.017 |
| | C | RadiologyNET | $10^{-4}$ | **0.695 ± 0.022** | **0.593 ± 0.026** |
| | N/A | Baseline | $10^{-4}$ | 0.688 ± 0.013 | 0.586 ± 0.014 |
| U-Net-ResNet50 | C | ImageNet | $10^{-4}$ | 0.692 ± 0.026 | 0.593 ± 0.03 |
| | C | RadiologyNET | $10^{-5}$ | **0.715 ± 0.017** | **0.616 ± 0.017** |
| | N/A | Baseline | $10^{-4}$ | 0.646 ± 0.027 | 0.538 ± 0.03 |
| U-Net-VGG16 | C | ImageNet | $10^{-5}$ | **0.729 ± 0.01** | **0.632 ± 0.015** |
| | C | RadiologyNET | $10^{-4}$ | 0.706 ± 0.015 | 0.605 ± 0.019 |
| | N/A | Baseline | $10^{-4}$ | 0.704 ± 0.03 | 0.601 ± 0.033 |

Among the results obtained for the basic U-Net (shown in Table 4.2), the *Baseline* model demonstrated superior performance compared to both TL strategies, achieving the highest Dice and IoU scores. In contrast, U-Net models pretrained on ImageNet and RadiologyNET using reconstruction tasks exhibited significantly lower performance (MWU, $p = 0.024$ for both ImageNet and RadiologyNET compared to *Baseline*). In comparison, U-Net-ResNet50, U-Net-EfficientNetB4, and U-Net-VGG16 demonstrated improved performance over the basic U-Net, obtaining higher Dice and IoU scores. Among these architectures, the performance of the three TL approaches was comparable, with the only statistically significant difference observed between the U-Net-ResNet50 model pretrained on RadiologyNET and the *Baseline* model (MWU, $p = 0.024$).

Figure 4.3 shows the difference in model outputs between reconstruction-pretrained models and classification-pretrained models. Reconstruction-pretrained models merely replicated the input image, indicating a lack of learned discriminative features despite attempts to impart valuable features to the U-Net models.

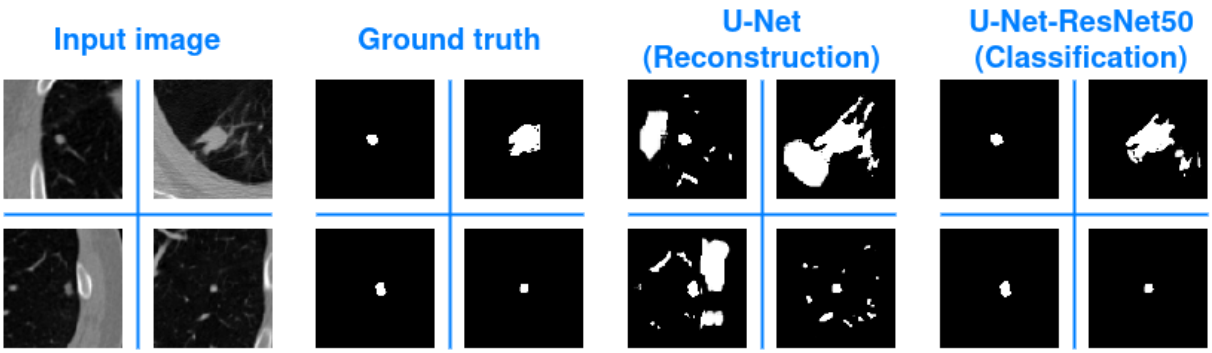



**Figure 4.3:** A figure showing the impact of different pretraining strategies of U-Net models on four randomly selected images from the LUNA dataset.

## Pediatric Bone Age Challenge Results

As shown in Table 4.3, the RadiologyNET-pretrained EfficientNetB3 model achieved the lowest MAE, outperforming the ImageNet-pretrained counterpart. However, ImageNet models exhibited significantly faster convergence (ImageNet vs. RadiologyNET, MWU, $p = 0.033$).

**Table 4.3:** Metric mean and standard deviation calculated on the test subset of Pediatric Bone Age Challenge, across five runs. Best results are emphasised.

| Challenge | TL Model | LR | RMSE | MAE | Epochs |
|---|---|---|---|---|---|
| PBA EfficientNetB3 (avg.) | ImageNet | $10^{-3}$ | 11.77 ± 1.4 | 9.37 ± 1.3 | **22.6 ± 7.4** |
| | RadiologyNET | $10^{-3}$ | **10.8 ± 0.3** | **8.23 ± 0.2** | 41.2 ± 1.1 |
| | Baseline | $10^{-3}$ | 31.98 ± 17.2 | 25.2 ± 13.9 | 30.8 ± 9.7 |
| PBA InceptionV3 (avg.) | ImageNet | $10^{-4}$ | **11.56 ± 1.1** | **9.08 ± 1.0** | **28.0 ± 8.2** |
| | RadiologyNET | $10^{-2}$ | 12.17 ± 0.4 | 9.31 ± 0.3 | 46.6 ± 9.9 |
| | Baseline | $10^{-3}$ | 12.16 ± 0.2 | 9.36 ± 0.3 | 45.4 ± 9.1 |
| PBA EfficientNetB3 (best) | ImageNet | $10^{-3}$ | **9.91** | **7.56** | **32.0** |
| | RadiologyNET | $10^{-3}$ | 10.971 | 8.261 | 41.0 |
| | Baseline | $10^{-3}$ | 12.572 | 9.2 | 41.0 |
| PBA InceptionV3 (best) | ImageNet | $10^{-4}$ | **11.154** | **8.587** | **41.0** |
| | RadiologyNET | $10^{-2}$ | 12.126 | 9.086 | 51.0 |
| | Baseline | $10^{-3}$ | 12.028 | 9.296 | 55.0 |

For the InceptionV3 architecture, ImageNet-pretrained models attained lower MAE compared to both RadiologyNET and *Baseline* models, but the results between the approaches were not statistically significant (Kruskal-Wallis, $p = 0.185$). ImageNet-pretrained models exhibited faster convergence than those pretrained on RadiologyNET, although this difference was not statistically significant (ImageNet vs. RadiologyNET, MWU, $p = 0.139$). RadiologyNET-pretrained InceptionV3 models demonstrated com-

parable performance to *Baseline* models in terms of convergence time (RadiologyNET vs. *Baseline*, MWU, $p = 1.00$).

## GRAZPEDWRI-DX Results

**Table 4.4:** Metric mean and standard deviation calculated on the test subset of GRAZPEDWRI-DX, across five runs. Best results are emphasised.

| Challenge | TL Model | LR | Acc (%) | F1-Score (%) | Epoch |
|---|---|---|---|---|---|
| GRAZPEDWRI DenseNet121 (avg.) | ImageNet | $10^{-3}$ | **93.1 ± 1.0** | **93.1 ± 1.0** | 24.6 ± 9.0 |
| | RadiologyNET | $10^{-4}$ | 92.0 ± 0.8 | 92.0 ± 0.8 | **15.2 ± 1.1** |
| | Baseline | $10^{-3}$ | 90.6 ± 2.4 | 90.6 ± 2.4 | 38.6 ± 10.4 |
| GRAZPEDWRI ResNet34 (avg.) | ImageNet | $10^{-3}$ | **92.6 ± 0.3** | **92.6 ± 0.3** | 24.4 ± 5.3 |
| | RadiologyNET | $10^{-3}$ | 91.5 ± 1.1 | 91.5 ± 1.0 | **20.4 ± 2.1** |
| | Baseline | $10^{-2}$ | 81.5 ± 11.6 | 80.4 ± 13.6 | 33.8 ± 14.4 |
| GRAZPEDWRI DenseNet121 (best) | ImageNet | $10^{-3}$ | 92.6 | 92.6 | 17.0 |
| | RadiologyNET | $10^{-4}$ | 92.9 | 92.9 | **16.0** |
| | Baseline | $10^{-3}$ | **93.2** | **93.2** | 51.0 |
| GRAZPEDWRI ResNet34 (best) | ImageNet | $10^{-3}$ | 92.4 | 92.4 | 28.0 |
| | RadiologyNET | $10^{-3}$ | **92.9** | **92.9** | **22.0** |
| | Baseline | $10^{-2}$ | 91.0 | 91.0 | 51.0 |

*Acc – Accuracy*

The results for GRAZPEDWRI-DX are given in Table 4.4. For the DenseNet121 architecture, ImageNet-pretrained models attained higher average F1-scores compared to other approaches; however, the differences in scores between the three methods were not statistically significant (Kruskal-Wallis, $p = 0.063$). RadiologyNET-pretrained models demonstrated the fastest convergence which, while not significantly different from ImageNet (ImageNet vs. RadiologyNET, MWU, $p = 0.07$), was significantly faster than the *Baseline* models (RadiologyNET vs.*Baseline*, MWU, $p = 0.033$).

The performance of *Baseline* ResNet34 models diverged between runs. When comparing F1-scores, ImageNet-pretrained models significantly outperformed *Baseline* models (ImageNet vs. *Baseline*, MWU, $p = 0.024$), while the difference between RadiologyNET and *Baseline* was less pronounced and not statistically significant (RadiologyNET vs. *Baseline*, MWU, $p = 0.095$). In terms of convergence, no statistically significant differences were observed between the approaches (Kruskal-Wallis, $p = 0.199$).

## COVID-19 Results

**Table 4.5:** Metric mean and standard deviation calculated on the test subset of COVID-19, across five runs. Best results are emphasised.

| Challenge | TL Model | LR | Acc (%) | F1-Score (%) | Epoch |
|---|---|---|---|---|---|
| COVID-19 MobileNetV3Large (avg.) | ImageNet | $10^{-3}$ | 97.1 ± 1.0 | 97.1 ± 1.0 | **23.6 ± 13.5** |
| | RadiologyNET | $10^{-4}$ | **97.7 ± 0.1** | **97.8 ± 0.1** | 25.4 ± 5.9 |
| | Baseline | $10^{-4}$ | 94.5 ± 1.6 | 94.5 ± 1.6 | 32.0 ± 5.5 |
| COVID-19 ResNet18 (avg.) | ImageNet | $10^{-4}$ | 97.9 ± 0.7 | **98.0 ± 0.7** | **16.0 ± 0.0** |
| | RadiologyNET | $10^{-4}$ | **98.0 ± 0.1** | **98.0 ± 0.1** | 26.6 ± 6.1 |
| | Baseline | $10^{-3}$ | 96.5 ± 0.5 | 96.5 ± 0.5 | 39.8 ± 6.9 |
| COVID-19 MobileNetV3Large (best) | ImageNet | $10^{-3}$ | 97.5 | 97.5 | 45.0 |
| | RadiologyNET | $10^{-4}$ | **97.8** | **97.8** | **23.0** |
| | Baseline | $10^{-4}$ | 96.0 | 96.0 | 40.0 |
| COVID-19 ResNet18 (best) | ImageNet | $10^{-4}$ | 97.3 | 97.3 | **16.0** |
| | RadiologyNET | $10^{-4}$ | **98.2** | **98.2** | 26.0 |
| | Baseline | $10^{-3}$ | 96.5 | 96.5 | 47.0 |

*Acc – Accuracy*

While CR images represent a minority within the RadiologyNET dataset, chest radiographs constitute the most prevalent subtype (Figure 3.3b). As a result, ImageNet and RadiologyNET models demonstrated comparable performance on MobileNetV3Large, with no statistically significant differences observed (ImageNet vs. RadiologyNET, MWU, $p = 1.00$). In contrast, models trained from scratch (*Baseline*) consistently underperformed compared to both ImageNet and RadiologyNET, achieving significantly lower F1-scores on the test subset (MWU, $p = 0.047$ and $p = 0.024$ for ImageNet and RadiologyNET, respectively). The number of epochs required for convergence did not differ significantly between the approaches when using MobileNetV3Large (Kruskal-Wallis, $p = 0.326$).

Regarding the evaluation of ResNet18 models, ImageNet and RadiologyNET pretrained models exhibited nearly identical F1-scores, with no statistically significant difference between them (ImageNet vs. RadiologyNET, MWU, $p = 1.00$). Both TL approaches significantly outperformed the *Baseline* models (MWU, $p = 0.035$ for both ImageNet and RadiologyNET). Notably, all ImageNet models converged consistently at the 16th epoch, which was significantly faster than both RadiologyNET and *Baseline* models (MWU, $p = 0.020$ and $p = 0.022$, respectively). The difference in convergence time between RadiologyNET and *Baseline* models was not statistically significant (Radi-

ologyNET vs. *Baseline*, MWU, $p = 0.103$).

## Brain Tumor MRI Results

**Table 4.6:** Metric mean and standard deviation calculated on the test subset of Brain Tumor MRI, across five runs. Best results are emphasised.

| Challenge | TL Model | LR | Acc (%) | F1-Score (%) | Epoch |
|---|---|---|---|---|---|
| BTMR MobileNetV3Small (avg.) | ImageNet | $10^{-4}$ | **98.1 ± 0.4** | **98.1 ± 0.4** | 41.4 ± 7.2 |
| | RadiologyNET | $10^{-4}$ | 97.9 ± 0.3 | 97.9 ± 0.3 | **39.4 ± 5.7** |
| | Baseline | $10^{-4}$ | 95.3 ± 2.3 | 95.1 ± 2.4 | 60.0 ± 17.8 |
| BTMR ResNet50 (avg.) | ImageNet | $10^{-5}$ | 98.7 ± 0.1 | 98.7 ± 0.1 | 41.6 ± 8.0 |
| | RadiologyNET | $10^{-4}$ | **98.9 ± 0.4** | **98.9 ± 0.4** | **21.2 ± 3.0** |
| | Baseline | $10^{-4}$ | 97.5 ± 0.8 | 97.4 ± 0.8 | 44.4 ± 10.7 |
| BTMR MobileNetV3Small (best) | ImageNet | $10^{-4}$ | 97.6 | 97.4 | **46.0** |
| | RadiologyNET | $10^{-4}$ | **98.0** | **98.0** | **46.0** |
| | Baseline | $10^{-4}$ | 97.6 | 97.6 | 83.0 |
| BTMR ResNet50 (best) | ImageNet | $10^{-5}$ | 98.6 | 98.6 | 47.0 |
| | RadiologyNET | $10^{-4}$ | **99.2** | **99.2** | **23.0** |
| | Baseline | $10^{-4}$ | 98.2 | 98.2 | 51.0 |

*Acc – Accuracy*

Out of all the downstream tasks, the Brain Tumor MRI dataset exhibits the highest overlap with the original pretraining dataset, as it comprises MR images of the brain (which are prevalent in the RadiologyNET dataset). The results for the BTMR dataset are presented in Table 4.6. For the MobileNetV3Small architecture, ImageNet and RadiologyNET models achieved nearly identical F1-scores (ImageNet vs. RadiologyNET, MWU, $p = 1.00$). In contrast, *Baseline* models displayed significantly lower classification metrics compared to both TL approaches (MWU, $p = 0.048$).

Similarly, for the ResNet50 architecture, ImageNet and RadiologyNET models demonstrated comparable performance (MWU, $p = 1.00$), while *Baseline* models performed significantly worse (MWU, $p = 0.024$ and $p = 0.036$ for ImageNet and RadiologyNET, respectively). However, RadiologyNET-pretrained models exhibited an advantage in convergence time, requiring significantly fewer epochs compared to *Baseline* (RadiologyNET vs. *Baseline*, MWU, $p = 0.028$). The difference in convergence time between ImageNet and RadiologyNET was insignificant (ImageNet vs. RadiologyNET, MWU, $p = 0.052$).

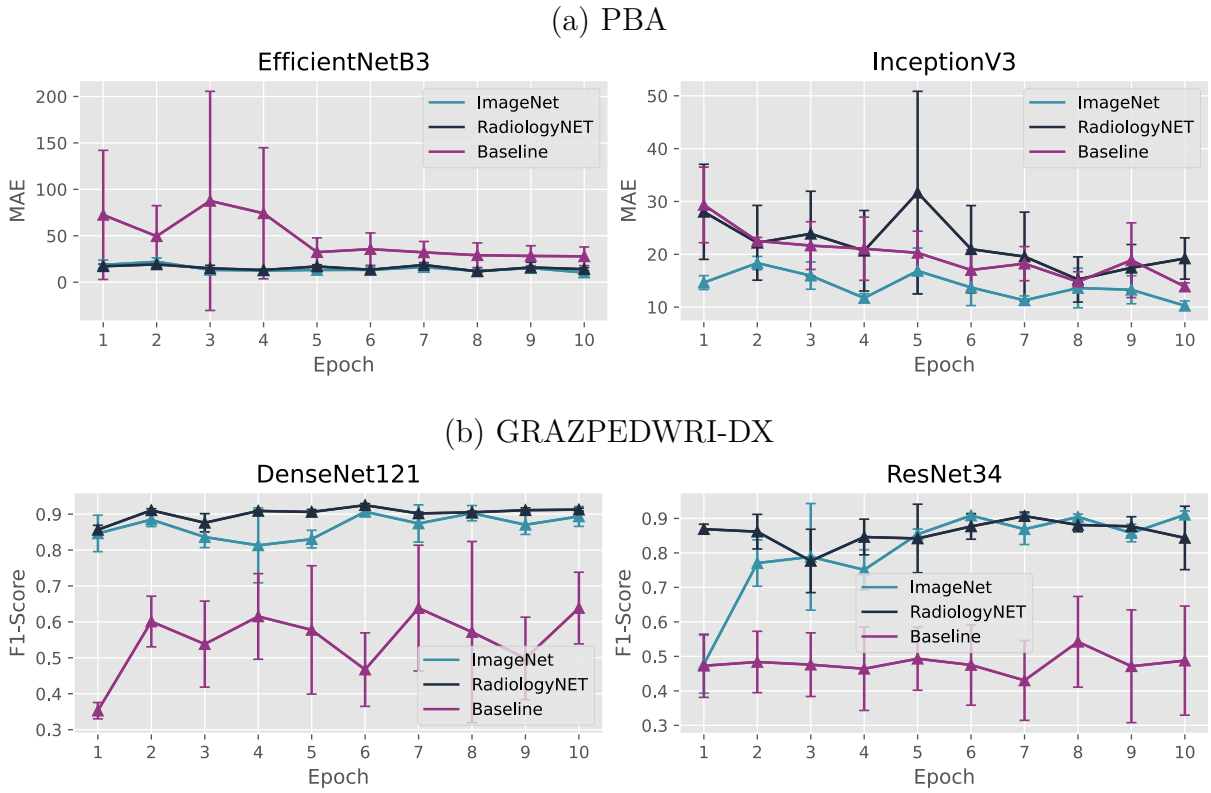



**Figure 4.4:** Average performance of best-performing models on the validation subset across the first 10 epochs on the PBA and GRAZPEDWRI-DX datasets. F1-score five-run mean and standard deviation is show per each epoch.

## 4.3. Training Progress and Resource-limited Conditions

Although the initial findings indicated minimal performance differences between the three approaches when models reached convergence, there were observed differences during the early stages of training (i.e. in the first few epochs, shown in Figures 4.4 and 4.5.). This observation led to the hypothesis that the three approaches might exhibit different behaviours when in resource-limited conditions, i.e. when training time and training data are significantly reduced. To investigate this potential difference and examine any advantages in initial model training, a small-scale experiment was conducted using the GRAZPEDWRI-DX and Brain Tumor MRI datasets. GRAZPEDWRI-DX was selected due to its minimal overlap with the RadiologyNET pretraining domain, while Brain Tumor MRI was chosen for its closer alignment with the pretraining data.

In this resource-limited experiment, training time was reduced to 10 epochs, while

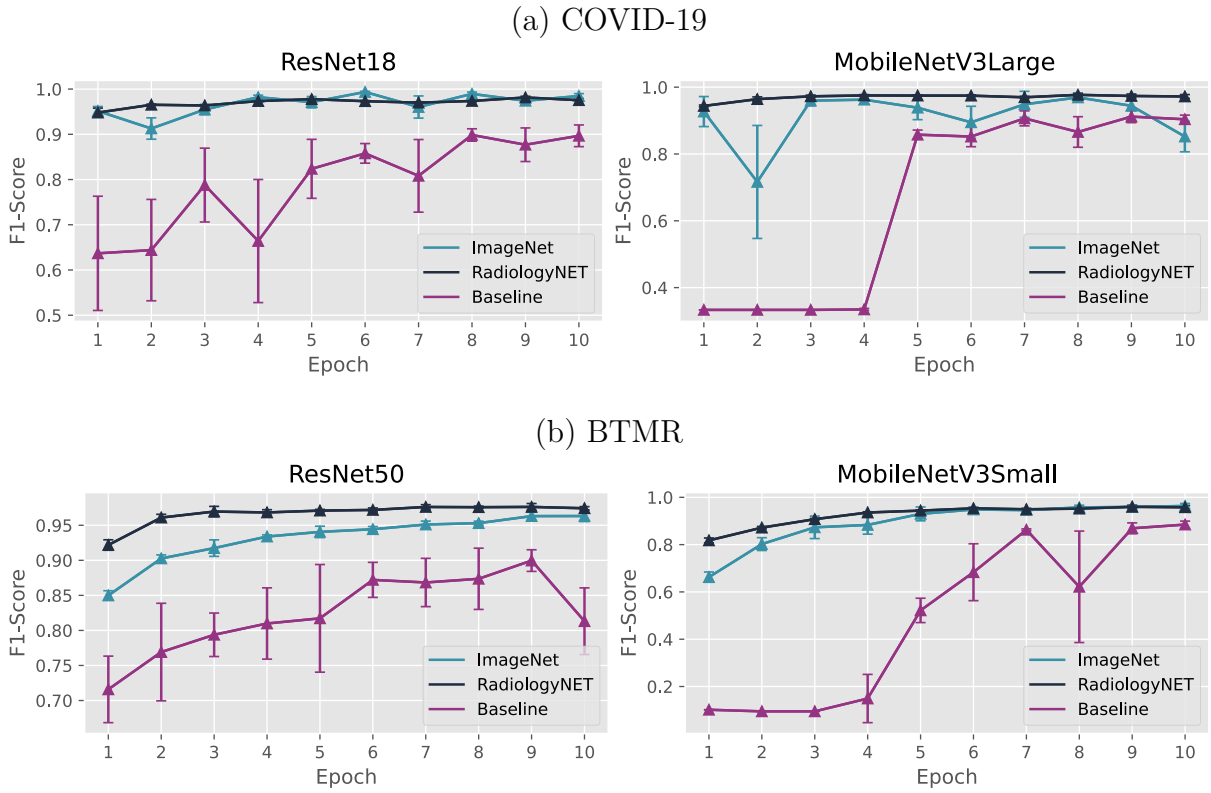



**Figure 4.5:** Average performance of best-performing models on the validation subset across the first 10 epochs on the COVID-19 and Brain Tumor MRI datasets. F1-score five-run mean and standard deviation is show per each epoch.

the training subsets were randomly undersampled to 5%, 25%, and 50% of their original size (Figure 4.6a). As is visible in the figure, only the training subset was changed, with the validation and test subsets remaining untouched. This undersampling process was performed in a way that does not change the distribution or overall quality of data, to avoid introducing any additional bias. An example distribution of undersampling at each level is shown in Figure 4.6b. Models were trained using the learning rates specified in Tables 4.4 and 4.6. Each approach was trained in five independent runs, and the mean F1-scores along with the standard deviation are presented in Figure 4.7.

## 4.4. Discussion

In most cases, RadiologyNET and ImageNet models exhibited similar performance, particularly when the training process was not constrained by data or time. Statistically significant differences were mainly observed between these two approaches against the *Baseline* models, with the latter generally demonstrating worse performance compared to

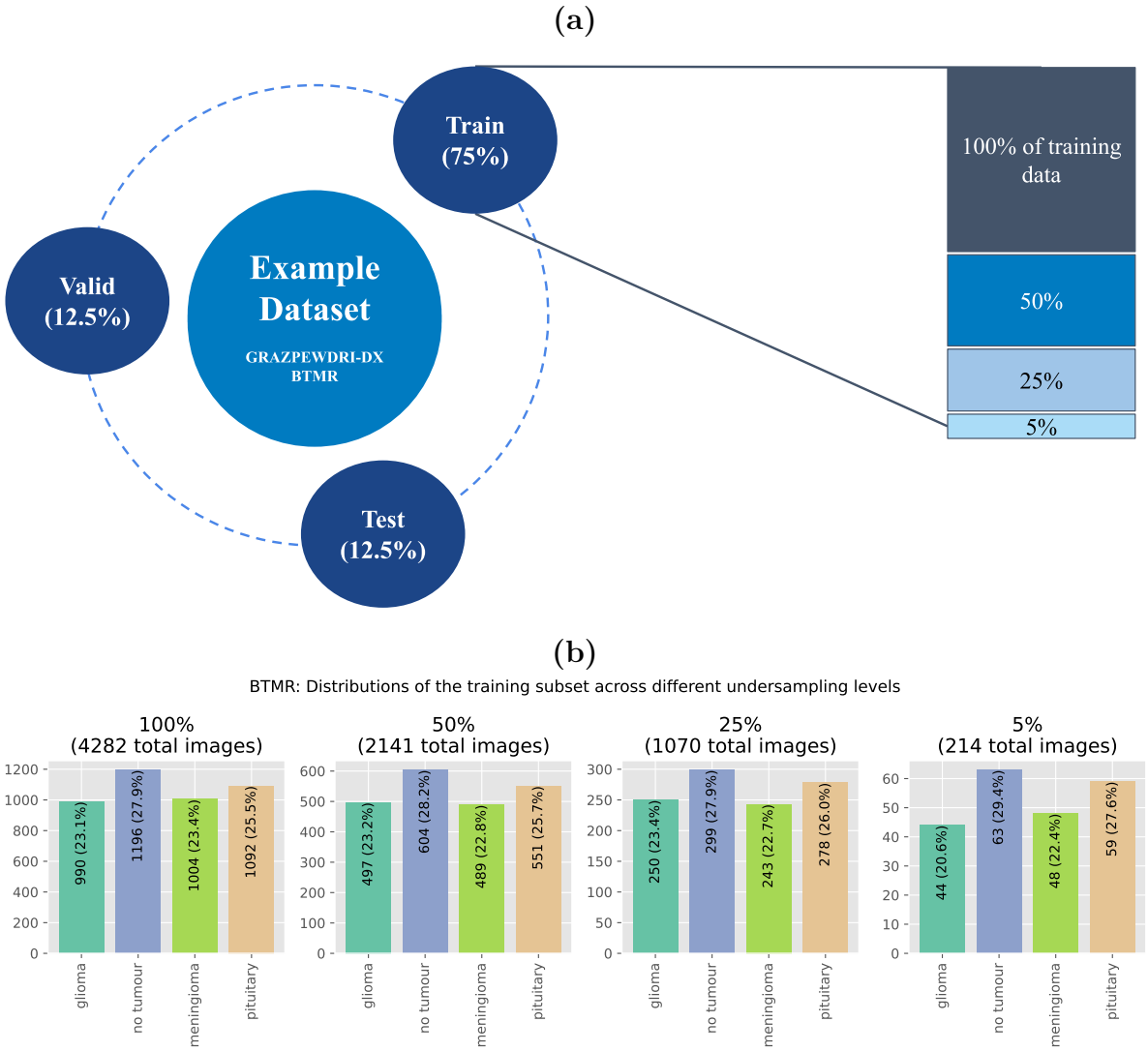
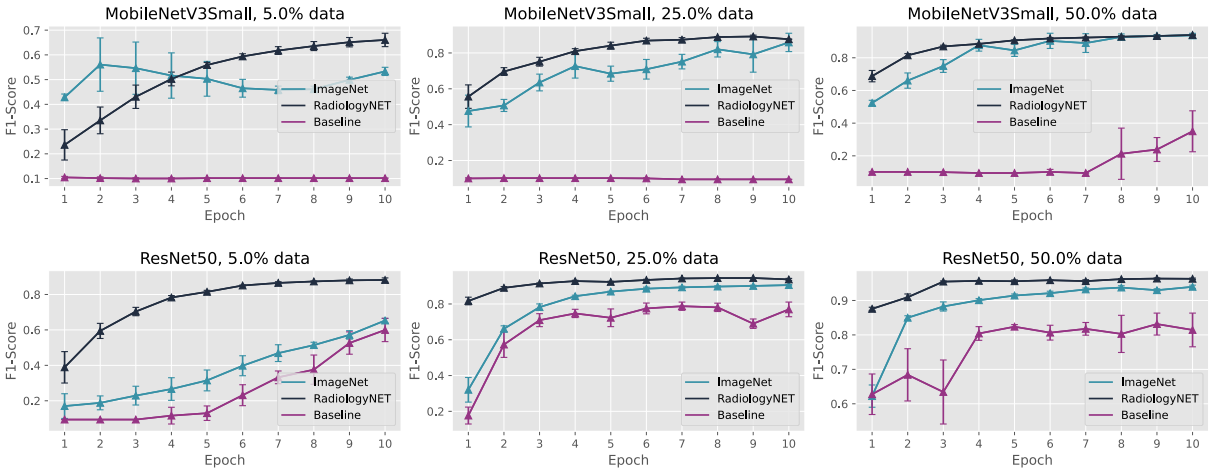



**Figure 4.6:** (a) Undersampling process performed on the training data, with the validation and test subsets remaining unchanged. (b) Class distribution at each level of undersampling on the BTMR dataset.

both ImageNet and RadiologyNET models.

When observing each challenge separately, an interesting pattern appeared in the LUNA dataset. Models pretrained as reconstruction tasks significantly underperformed compared to those pretrained as classification tasks. The reason behind this could be that reconstruction pretraining merely focused on replicating textures and patterns rather than capturing the semantic meaning of each pixel, resulting in outputs that only mimic the input image (Figure 4.3). In contrast, classification-pretrained encoders (ResNet50, VGG16, and EfficientNetB4) appeared to be better suited for segmentation tasks where pixel-wise semantic meaning is important, as is the case in LUNA nodule segmentation.

(a) BTMR




(b) GRAZPEDWRI-DX

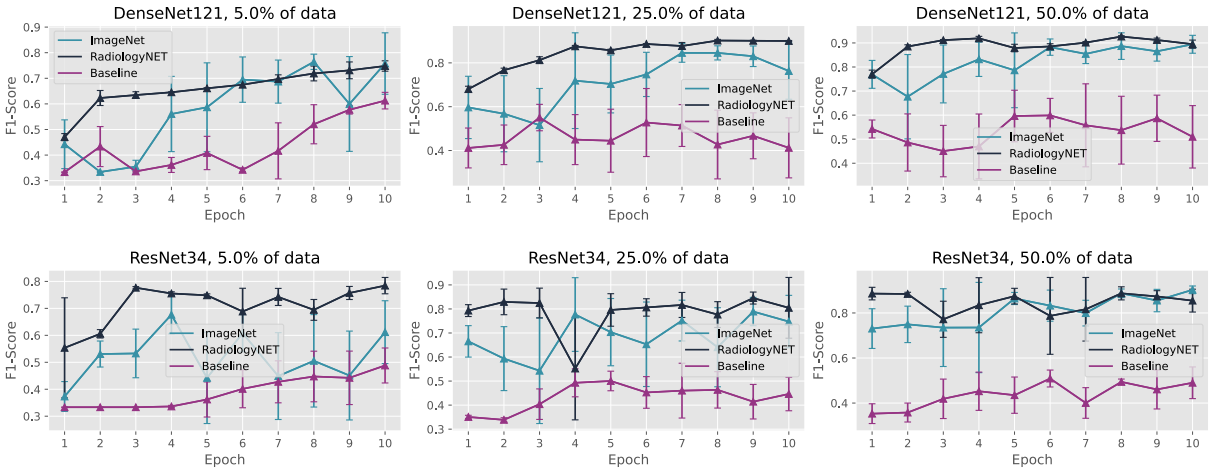



**Figure 4.7:** Results over 10 epochs for MobileNetV3Small and ResNet50 (on BTMR); and DenseNet121 and ResNet34 (on GRAZPEDWRI-DX), with training data reduced to 5%, 25%, and 50% of the original training set. The F1-score is reported as the mean and standard deviation across five runs for each epoch.

While reconstruction-pretrained models showed significantly worse performance, it is possible that they may perform better on other task types, such as image compression or denoising. However, evaluating this hypothesis was beyond the research scope in this context.

Other challenges also demonstrated the influence of the pretraining domain on TL efficacy. In the RSNA PBA and GRAZPEDWRI-DX tasks, the pretraining dataset (RadiologyNET) mainly consisted of CT and MR images of the head and abdomen, with limited representation of wrist and hand radiographs. Although ImageNet does not include medical images, its diverse range of natural images may have enabled ImageNet-pretrained models to learn more generalisable features compared to the more domain-specific RadiologyNET models. This observation was supported by the COVID-19 and BTMR results, where RadiologyNET models demonstrated comparable performance to ImageNet and, in some cases, better training progress compared to both ImageNet and *Baseline* models (Figure 4.5). Both ImageNet and RadiologyNET pretrained models demonstrated performance improvements within the first 10 epochs, particularly when compared to *Baseline*. The most notable boost from ImageNet was observed with the InceptionV3 architecture on the RSNA PBA Challenge, where it achieved a lower MAE than the other two approaches. On the other hand, RadiologyNET pretrained weights showed improved performance on DenseNet121, ResNet50, and MobileNetV3Small architectures, which is further supported by the overall reduction in the number of epochs required to reach convergence (as reported in Tables 4.4 and 4.6). An important thing to note is that the extent of this improvement may vary depending on the specific architecture and task, and it does not always result in statistically significant differences in final performance, which is a limitation of RadiologyNET models in their current form.

The greatest performance differences were observed under resource-limited conditions. As there were cases where *Baseline* models achieved comparable results when resources were not restricted, this indicates that the original challenges may have had sufficient training data, and that when the training pool is large enough, the advantages of TL become less impactful [101]. In Figure 4.7, it is clear that models where TL was applied show better performance against training from randomly initialised weights. Although RadiologyNET models did not outperform ImageNet in less-restricted resource conditions on the GRAZPEDWRI-DX dataset (i.e. the results shown in Table 4.4), they showed com-

petitive performance when training data and time were limited. However, it is important to note that as more training data becomes available (e.g. when the dataset is reduced to 50% instead of 5% of its original size), the performance differences between RadiologyNET and ImageNet become less pronounced. This suggests that the relative advantage of RadiologyNET pretraining may decrease as the availability of training data increases (as does the advantage of TL in general).

In the RadImageNet study [20], which evaluated TL on a dataset of comparable scale to RadiologyNET but annotated by 20 expert radiologists, statistically significant performance improvements were reported. The authors observed area-under-curve improvements of 1.9%, 6.1%, 1.7%, and 0.9% over ImageNet-pretrained models across five medical imaging tasks. While RadiologyNET-pretrained models demonstrated performance comparable to ImageNet in the experiments presented here, the results from RadImageNet underscore the benefits of expert supervision in improving TL efficacy. Although both datasets are similar in size, RadImageNet includes 165 pathology-specific labels, compared to the 36 pseudo-labels derived in RadiologyNET. This richer label space likely encourages models to learn more generalisable features, and features that are pathology-focused. Nevertheless, RadiologyNET achieved competitive performance without manual annotation, thereby avoiding the considerable annotation effort required for RadImageNet, which involved 20 radiologists.

The pretrained weights for all RadiologyNET models, along with code for TL, finetuning, and evaluation on downstream tasks, are publicly available at `https://github.com/AIlab-RITEH/RadiologyNET-TL-models`. This repository enables full reproducibility of the experiments described in this chapter.

**From the obtained results, we attempt to answer the following questions:**

1. **Does the domain of pretrained models affect performance?**

   Yes. RadiologyNET models demonstrated competitive performance; however, ImageNet-pretrained models exhibited a slight advantage on tasks such as the RSNA Pediatric Bone Age Challenge and the GRAZPEDWRI-DX dataset. This is can be attributed to (i) the greater diversity of images in ImageNet (which may contribute

to more generalisable feature representations), as well as (ii) the limited presence of wrist radiographs in the RadiologyNET dataset. In contrast, the Brain Tumor MRI dataset showed that RadiologyNET-pretrained models exhibited better training progress (and faster convergence), likely due to a better alignment between the pretraining domain and the downstream task. These results suggest that when selecting pretrained models for medical machine learning tasks, it is important to account for potential domain biases in the pretraining dataset (e.g. the distribution of anatomical regions in the pretraining data).

2. **Does the pretraining task matter?**

   Yes, the choice of pretraining task may play a key role. Selecting an appropriate pretraining task should align with the types of features relevant to the downstream task. Results from the LUNA Challenge demonstrated that models pretrained on classification objectives significantly outperformed those pretrained on image reconstruction when applied to semantic segmentation. This suggests that reconstruction-based pretraining may not be the best for segmentation tasks, likely due to the fundamentally different features learned. The ability to learn generalisable representations from diverse tasks is one of the foundations of multi-task meta-learning, in which models *learn to learn*. This approach, alongside multi-task learning, is the underlying principle for many foundation models [102].

3. **Is TL always beneficial?**

   The results indicate that TL generally improves model performance, especially under resource-constrained conditions (limited training data and training time). However, in some cases, *Baseline* models (trained from randomly initialised weights) achieved performance comparable to pretrained counterparts, and there is evidence to show that when sufficient data and training time are available, the benefits of TL may decrease [16, 101]. Notably, in the case of the reconstruction-pretrained U-Net, *Baseline* models significantly outperformed both ImageNet and RadiologyNET-pretrained variants, indicating that in some settings, TL may even hinder performance. These observations relate closely to the importance of pretraining task and domain alignment. The effectiveness of TL is dependent not only on the generalisability of learned features but also on how well the pretraining configuration

matches the requirements of the downstream task.

4. **What should be considered when collecting data for pretraining medical models for TL?**

   The results suggest that pretraining datasets with higher variability offer greater utility than homogeneous datasets. This observation aligns with trends in natural language processing, where general-domain corpora have been used to improve domain-specific performance. For instance, GatorTron [24] incorporated data from sources such as Wikipedia, and Med-PaLM [102] was developed by adapting general-purpose language models to the medical domain. In the same manner, the integration of diverse data (e.g. non-medical sources such as ImageNet) may contribute positively to the development of medical foundation models. This raises an important direction for future research: how can diverse data sources, both intra- and cross-domain, be combined to maximise transferability in medical TL?

# 5. Chapter

# DOMAIN INFLUENCE ON PERFORMANCE AND INTERPRETABILITY

In the previous chapter, it was observed that the pretraining dataset can have a significant impact downstream performance. Specifically, models pretrained on RadiologyNET demonstrated better performance on tasks where the domain of RadiologyNET aligned well with the target data (e.g. Brain Tumor MRI), rather than tasks where the domains did not align (such as GRAZPEDWRI-DX). Models pretrained on ImageNet exhibited a more generalisable performance across diverse downstream tasks, likely due to the diverse nature of ImageNet (which consists of millions of natural images). Therefore, it is possible that ImageNet's generalisation capabilities emerge from its exposure to diverse data, while RadiologyNET's domain-specific features can provide an advantage only when the downstream task aligns with the pretraining data.

From these observations, it is clear that patterns learned during pretraining influence downstream task performance, but the extent of this influence can be tested further. Two questions arise:

1. **Do the patterns learned during pretraining influence model interpretability?** For example, do these learned patterns affect which neurons are activated during inference on downstream tasks?

2. **Would TL models benefit more from being pretrained on modality-specific**

**data** (e.g. CR-only data for CR-based tasks), or would it be more beneficial to use models pretrained on diverse, multi-modality datasets?

At the time of writing this thesis, there is no universally accepted method for objectively evaluating model interpretability, and neural networks remain considered *black boxes*. In the context of image analysis, one of the most prevalent techniques is to examine activation maps through visualisation methods such as Gradient-weighted Class Activation Mapping (Grad-CAM) [103]. This approach highlights areas within the input image that activate specific neurons. Despite this approach not being perfect (being called *false hope* by Ghassemi et al. [104]), it is still one of the most commonly used approaches to examine model interpretability [93].

The second question can more easily be answered by pretraining several models on single-modality data and subsequently benchmarking them against their multi-modality counterparts. This was already described in the previous chapters, with the exact lists of tested architectures and modality-specific pretraining being given in Tables 3.1 and 4.1.

## 5.1. Grad-CAM Evaluation

To assess model interpretability, Grad-CAM [103] heatmaps were generated to visualise the areas of focus for each of the three approaches on the GRAZPEDWRI-DX and BTMR datasets. These two datasets were chosen due to their alignment with the RadiologyNET dataset, with BTMR being highly aligned and GRAZPEDWRI-DX being the opposite. In a Grad-CAM heatmap, warm colours indicate high activation, with red colour representing the highest value. Warm-coloured areas are the regions in the input image that contributed strongly to the model's decision or prediction. On the other hand, cool colours indicate low activation, with blue representing no recorded activation. Cool-coloured regions had less influence or were largely ignored by the model during its decision-making process.

The heatmaps were independently evaluated by two expert radiologists from different clinical centres (who are also situated in two different countries), to ensure the evaluation is as unbiased as possible. Each radiologist examined a sample of 20 randomly selected heatmaps from the GRAZPEDWRI-DX test set (obtained using DenseNet121) and another 20 heatmaps from the BTMR test set (generated using ResNet50). The evaluation

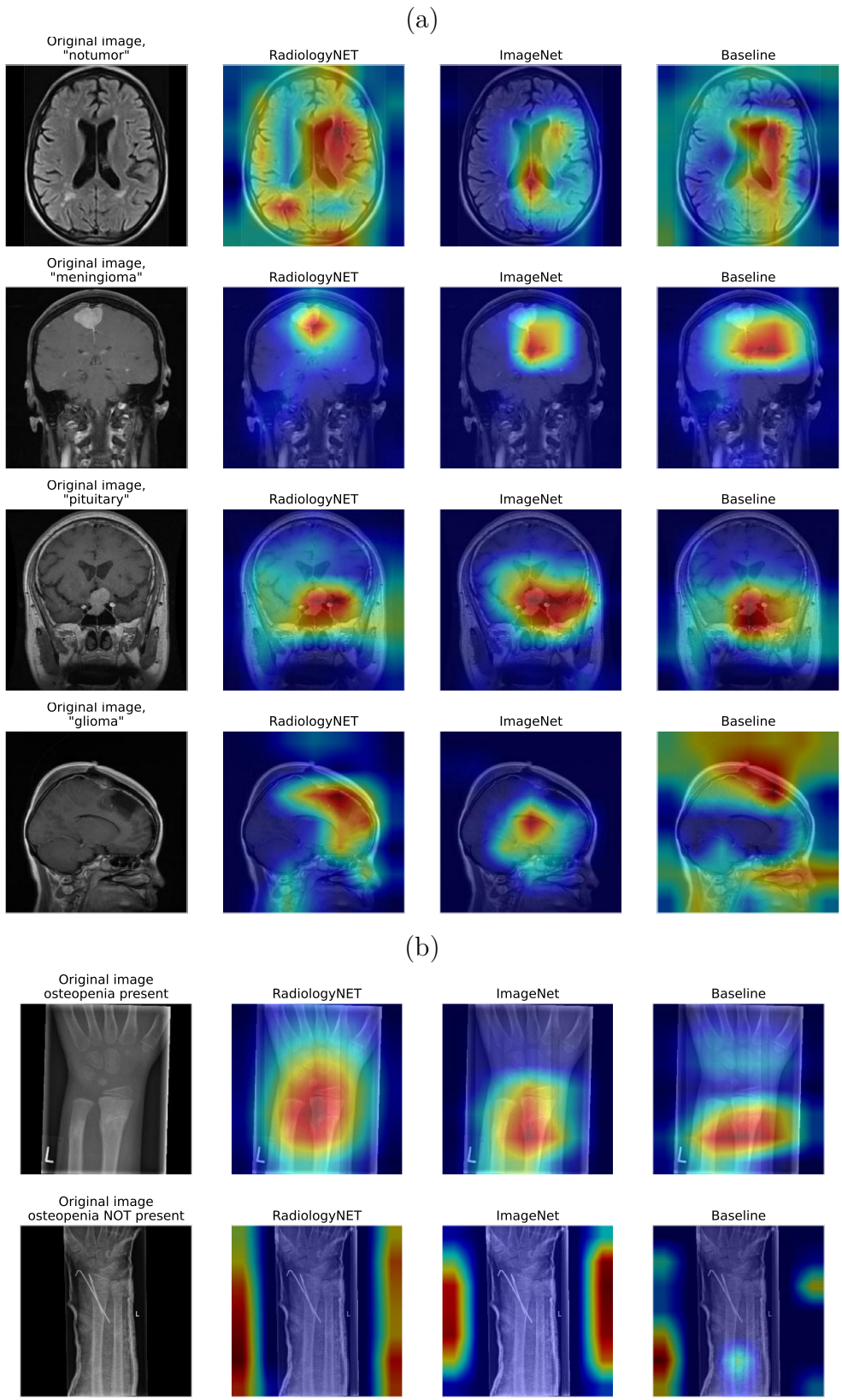



**Figure 5.1:** Grad-CAM heatmap examples of randomly selected images from: (a) the Brain Tumor MRI dataset (ResNet50); (b) GRAZPEDWRI-DX dataset (DenseNet121).


**Figure 5.2:** Ratings of each radiologist given to randomly sampled Grad-CAM heatmaps from the GRAZPEDWRI-DX and Brain Tumor MRI datasets.

involved rating each heatmap on a scale from 1 to 5, where 1 indicated that the model concentrated on entirely irrelevant areas, and 5 indicated focus exclusively on relevant regions. In addition to rating, the radiologists were encouraged to document any observations regarding the presented heatmaps. To eliminate potential bias, the source of each heatmap was concealed; they were labelled as algorithms *(a)* - RadiologyNET, *(b)* - ImageNet, and *(c)* - Baseline. This labelling ensured that the evaluation was based solely on the visual information presented.

Examples of the generated heatmaps shown to radiologists are provided in Figure 5.1. In this figure, there are four heatmaps presented for BTMR (one for each class, i.e. *no tumour*, *meningioma*, *pituitary* and *glioma*), and two heatmaps for GRAZPEDWRI-DX (one where osteopenia is present, and one where it is not). More heatmaps can be seen in Figures B.1, B.2, B.3, B.4, B.5, and B.6 which can be found in the Appendix. The radiologists' evaluation scores are presented in Figure 5.2.

Both radiologists noted that the BTMR heatmaps generated by *Baseline* models were unreliable, whereas RadiologyNET's heatmaps demonstrated the best focus on the pathologies present in the images. One radiologist observed that ImageNet's BTMR heatmaps appeared to be slightly offset in some instances, while the other noted that they were significantly less accurate in identifying tumour areas compared to RadiologyNET.

Regarding the GRAZPEDWRI-DX heatmaps, one radiologist reported that the presence of a cast puzzled all three models but noted that the *Baseline* model tended to *"focus a lot on the fracture [near osteopenia], but also the carpal bones, which would be the most relevant to look at."* The other radiologist stated that RadiologyNET's GRAZPEDWRI-DX heatmaps were generally the most reliable among the three, although they exhibited excessively wide areas of focus. ImageNet's heatmaps were described as inconsistent, sometimes displaying a high degree of specificity and accuracy, while at other times failing to capture the relevant area entirely. Both radiologists agreed that all three algorithms struggled with images where no pathologies were present.

To summarise, the radiologists' evaluation of the generated heatmaps indicated that RadiologyNET models were perceived as the most reliable overall, focusing on the present pathologies better than ImageNet and (especially) *Baseline*. This result raises questions about the influence of TL on model interpretability, as patterns learned during pretraining might help models focus on relevant regions in the downstream tasks. While pretraining on natural images can provide generalisable features, pretraining on medical data may lead to models that are better adapted to the specific characteristics of medical images (e.g. disease-related patterns and abnormalities). However, it is important to note the limitations of this experiment, as there were only two radiologists and two datasets with 20 randomly sampled images. To accurately confirm the extent of these findings, this experiment would have to be expanded upon by including more radiologists, a more diverse set of datasets – and therefore a greater number of sampled images.

## 5.2. Multi-modality versus Single-modality Pretraining

To evaluate the impact of modality-specific pretraining, an additional set of experiments was conducted, where the goal was to compare the performance of models pretrained on single-modality data (MR-only, CR-only, and CT-only) against their multi-modality pretrained counterparts. The evaluation was performed on the BTMR, RSNA PBA, COVID-19, GRAZPEDWRI-DX, and LUNA datasets.

For the segmentation task in the LUNA dataset, a ResNet50 model was pretrained

exclusively on CT images and integrated as the encoder in a U-Net-ResNet50 architecture. For the BTMR classification task, MR-only pretraining was performed using MobileNetV3Small and ResNet50 models. For the GRAZPEDWRI-DX, RSNA PBA, and COVID-19 downstream tasks, CR-only pretraining was applied to multiple architectures, including DenseNet121, ResNet34, EfficientNetB3, InceptionV3, MobileNetV3Large, and ResNet18. This goal was to determine whether modality-specific pretraining offers a performance advantage, particularly when the downstream task aligns with the modality used during pretraining.

Figure 5.3 illustrates the performance comparison between single-modality (MR-only, CR-only, and CT-only) and multi-modality pretrained RadiologyNET models. The results are presented across five independent runs, with statistical significance evaluated using Independent samples t-test. The results for all tested datasets and single-modality pretrained models are presented in the following subfigures: PBA (subfigures *a* and *b*), LUNA (subfigure *c*), GRAZPEDWRI-DX (subfigures *d* and *e*), COVID-19 (subfigures *f* and *g*), and BTMR (subfigures *h* and *i*). Mean and standard deviation are shown on top of (or above – as with PBA InceptionV3) each bar. Pairwise *p*-values are shown atop each tested pair, with statistically insignificant results marked in gray, and statistically significant results in blue.

Among the tested datasets (and architectures), U-Net-ResNet50 (LUNA dataset) was the only case where no statistically significant differences were observed between single-modality and multi-modality pretrained models. On the other hand, the PBA dataset exhibited mixed results depending on the architecture: for EfficientNetB3, multi-modality pretraining demonstrated significantly better performance than CR-only at all learning rates; while the opposite is true in the case of InceptionV3. In the GRAZPEDWRI-DX dataset, DenseNet121 models pretrained with multi-modality data generally achieved a significantly higher F1-score compared to CR-only models, with similar performance at the highest tested learning rate, $10^{-3}$. In contrast, ResNet34 models showed comparable performance between multi-modality and CR-only pretraining, with CR-only demonstrating a slight advantage at the lowest tested learning rate, $10^{-5}$. In the COVID-19 dataset, both multi-modality pretrained architectures (MobileNetV3Large and ResNet18) either outperformed CR-only models, or showed comparable performance with no statistically difference. In the BTMR dataset, MobileNetV3Small models pretrained with

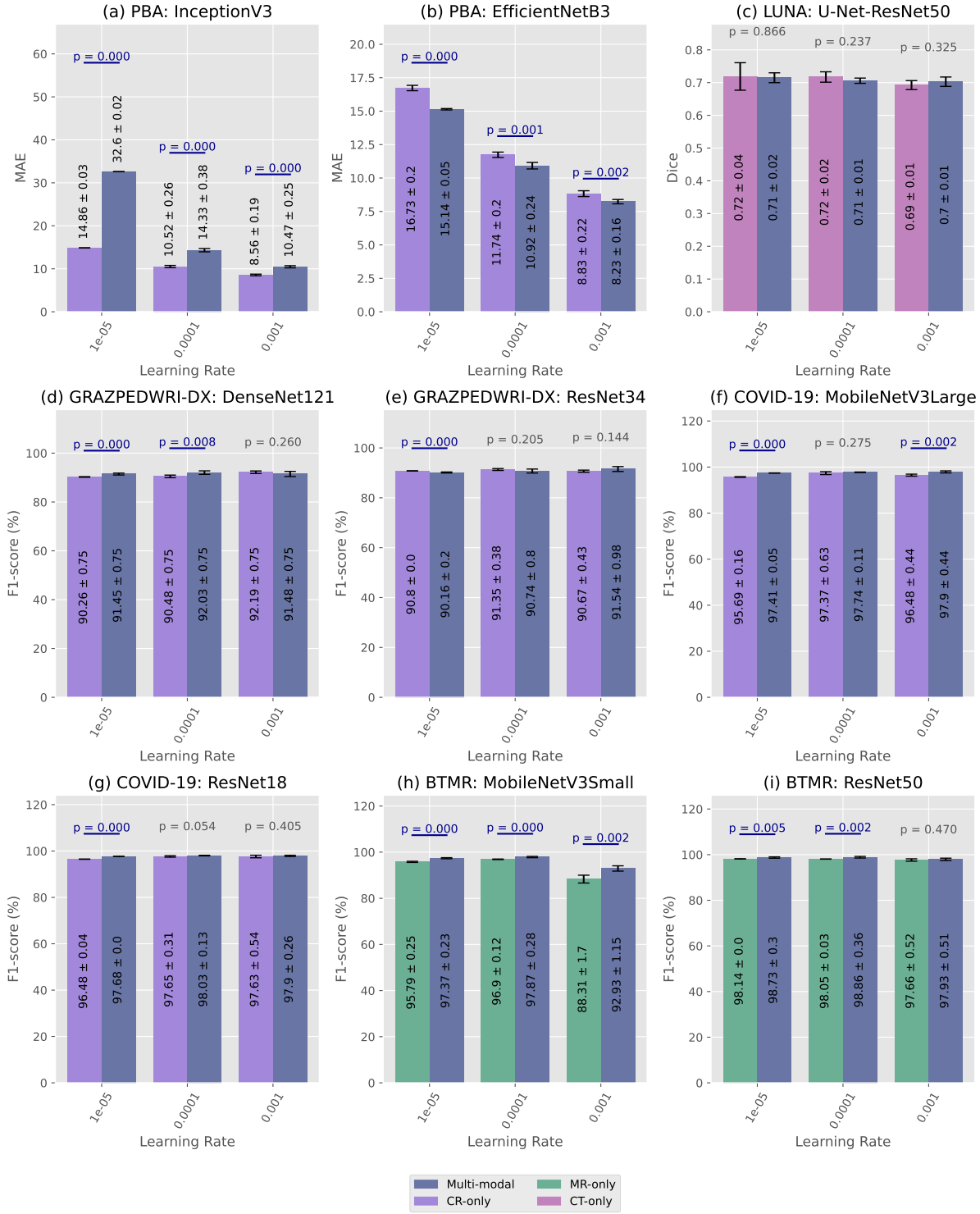



**Figure 5.3:** Comparison of TL performance across modality-specific (MR-only, CR-only, CT-only) and multi-modality pretrained RadiologyNET models.

multi-modality achieved significantly better performance compared to MR-only models across all tested learning rate settings. For ResNet50, multi-modality pretraining showed superior performance overall, although the differences were less pronounced at the highest tested learning rate, $10^{-3}$.

A total of 27 statistical comparisons were performed. Among these, 10 comparisons showed no statistically significant differences. In four instances, single-modality pretraining exhibited superior performance, while multi-modality pretraining demonstrated better results in 13 cases. In the MR-only comparisons, multi-modality pretraining outperformed MR-only in five out of six cases. For CR-only comparisons, single-modality pretraining demonstrated improved performance in four out of 18 cases, whereas multi-modality pretraining proved better in nine cases. In the CT-only comparisons on the LUNA dataset, no statistically significant differences were observed between CT-only and multi-modality pretraining.

For the BTMR classification task, models pretrained on MR-only data showed a statistically significant drop in performance compared to their multi-modality counterparts, suggesting that additional modality diversity enables the model to learn a broader and more transferable set of visual patterns. A similar trend was observable in the COVID-19 dataset, where multi-modality pretrained models (MobileNetV3Large and ResNet18) generally outperformed CR-only models or showed comparable performance with no significant differences. In the RSNA PBA dataset, the results were architecture-dependent: for EfficientNetB3, multi-modality pretraining demonstrated significantly better performance than CR-only across all learning rates, while for InceptionV3, the opposite was true, with CR-only pretraining consistently outperforming multi-modality. In GRAZPEDWRI-DX, DenseNet121 generally exhibited better multi-modality performance, and ResNet34 showed mixed results. However, for the LUNA segmentation task, where CT is the most prevalent modality in RadiologyNET (comprising 53.73% of the dataset), no significant performance difference was observed between CT-only and multi-modality pretrained models.

The results indicate that the choice of neural network topologies (and their internal mechanisms) could be a factor, but there is also a general trend visible: modality diversity is especially valuable when the single modality lacks sufficient internal variability or representation. For example, CR-only models were pretrained on X-ray images, which,

despite having less overall training data, exhibit high internal variability and do not necessarily benefit from other modalities or anatomical regions (e.g. patterns found in lung CT images do not consistently contribute to skeletal age prediction from hand CR images). On the other hand, MR-only models did not perform as well as multi-modality models on the BTMR task, suggesting that patterns found in CT images of the brain may contribute towards brain tumour classification in MR images. Either way, this further contributes to the idea that diverse pretraining data improves TL generalisability; i.e. using heterogeneous data did not consistently benefit from incorporating data that is inherently homogeneous, but homogeneous data did benefit from incorporating other modalities. This brings another question into focus: is it better to pretrain on multiple anatomical regions captured in the same modality, or is it better to pretrain on a single anatomical region captured in different modalities? However, testing this question fell out of scope of the research presented here.

# 6. Chapter

# CONCLUSION

This thesis explored the construction and evaluation of domain-specific foundation models for medical image analysis, with a focus on TL. The key contributions of this research were to develop a method for automated grouping and pseudo-labelling semantically similar medical radiology images; and to develop and evaluate RadiologyNET foundation models for TL in medical images. The work presented here contributes new insights into data curation and model pretraining.

A dataset of 25 million radiology images acquired from Clinical Hospital Centre Rijeka was collected, totalling 13 terabytes and comprising three data types: images, metadata and narrative diagnoses. A query-capable framework was implemented for fast data point retrieval and filtering. Through examining the query-capable database, it was concluded that none of the data types carried sufficient information to be used as *de-facto* labels, leading to the hypothesis that *feature extraction* could be the path forward – and that patterns within data points could be used to group semantically similar data points together. To this end, multiple feature extractors were used on each data type, with an extensive ablation study performed for each of them. This process resulted in 36 pseudo-labels which would be used for supervised model pretraining without requiring expert labelling.

Multiple CNN architectures were pretrained on the constructed pseudo-labelled RadiologyNET dataset. The pretrained architectures included U-Net, InceptionV3, DenseNet121, and multiple topologies from the ResNet, MobileNet and EfficientNet families. The pretraining was performed as a multi-class classification task using the pseudo-labels generated in the first phase, with additional experiments involving reconstruction-based

pretraining for the U-Net architecture. To this end, the entire ImageNet dataset was downloaded and pretrained as a reconstruciton task using the U-Net model. Both multi-modality and modality-specific subsets of the data were used to analyse the impact of pretraining diversity.

The pretrained RadiologyNET models were evaluated across five publicly available downstream medical datasets, covering a variety of imaging modalities (CR, CT, MR), anatomical regions (lungs, chest, head, wrists/hands) and task types (segmentation, binary and multiclass classification, and regression). The three TL strategies compared were: (i) training from randomly initialised weights (*Baseline*), (ii) fine-tuning from ImageNet, and (iii) fine-tuning from RadiologyNET. The results showed that RadiologyNET models performed comparably to ImageNet models overall, and offered a performance advantage in resource-limited conditions (e.g. limited training data or reduced training time), but it is also important to acknowledge that RadiologyNET did not consistently outperform ImageNet, which is a limitation of RadiologyNET models in their current form. This is particularly evident in downstream tasks where the target domain did not align with the pretraining domain, such as wrist radiographs (which were scarce in the pretraining data). Despite this, RadiologyNET models showed improved performance when the downstream task aligned with the pretraining data, such as brain tumour classification in MR images.

The complete codebase for performing TL, fine-tuning, and evaluation on downstream tasks, as well as the pretrained RadiologyNET weights, have been made publicly available at `https://github.com/AIlab-RITEH/RadiologyNET-TL-models`. The goal of releasing these pretrained models and the codebase publicly is to assist in the progress of medical machine learning and computer-aided diagnosis systems. These weights provide the basis for our future research in the field of medical TL; by iteratively improving the pretraining task and/or pretraining data, future versions of RadiologyNET models can be compared against the baseline provided here.

The main advantage ImageNet has over RadiologyNET is the diversity of pretraining data. While ImageNet does not contain medical radiology images, having diversity in pretraining data leads to improved generalisation capabilities. Therefore, RadiologyNET models did not outperform ImageNet in downstream tasks which did not align with RadiologyNET's domain, as the RadiologyNET dataset is relatively homogeneous

considering the capturing modality and the observed anatomical regions. To further test this, multiple experiments were conducted on single and multi-modality pretrained models, which confirmed this: using heterogeneous data where the intra-domain variability is high did not consistently benefit from incorporating data that is inherently homogeneous, but homogeneous data did benefit from incorporating other modalities.

Nonetheless, several limitations must be acknowledged. First, the unsupervised annotation approach relied on clustering and no pathology labels were included (in contrast, RadImageNet is labelled using pathology-oriented labels). Second, the dataset was derived from a single clinical centre and, as standard practice may differ between hospitals, this may limit generalisability. Third, although RadiologyNET models demonstrated advantages in training efficiency, they did not consistently outperform ImageNet in final performance metrics. This suggests that further refinement is required to fully realise the potential of RadiologyNET pretrained models. Nonetheless, labelling through a fully unsupervised approach and then pretraining on the generated pseudo-labels did consistently outperform training from randomly initialised weights, which suggests that the generated pseudo-labels may be a good starting point for further research. Moreover, the initial phases of this research were limited by the available hardware, which led to the usage of simpler representation learning methods such as autoencoders and clustering. In future iterations of RadiologyNET foundation models, more advanced unsupervised or self-supervised annotation/pretraining could be tested, such as contrastive learning via CLIP [19, 92].

In conclusion, the results presented in this thesis support the two central hypotheses: (i) that unsupervised annotation of medical data is a viable method of labelling medical data and grouping semantically similar data points together, facilitating the dataset's use for building foundation models, and (ii) that TL from RadiologyNET foundation models can improve model performance, especially in resource-limited medical tasks. First, the findings confirm that unsupervised annotation through clustering of semantically similar data points based on multimodal features offers a viable strategy for constructing large-scale labelled datasets in the absence of expert annotation. Pretraining from pseudo-labels consistently outperformed randomly initialised models across a variety of tasks. Expert annotation still remains the gold standard, but this approach to unsupervised labelling can serve as a first step, and possibly be the basis towards more fine-grained expert labels.

Second, the results show that transfer learning from RadiologyNET-pretrained models improves performance under resource-constrained conditions (i.e. limited training data or training time). While RadiologyNET models did not outperform ImageNet-pretrained counterparts in final performance metrics, they provided significant advantages when there was domain alignment and in early training phases. These findings underscore the importance of pretraining data diversity, domain alignment, and task-specific considerations in the development of medical foundation models.

# BIBLIOGRAPHY

[1] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "Ai in health and medicine," *Nature Medicine*, vol. 28, no. 1, pp. 31–38, Jan 2022. [Online]. Available: https://doi.org/10.1038/s41591-021-01614-0

[2] P. Oza, P. Sharma, and S. Patel, *Machine Learning Applications for Computer-Aided Medical Diagnostics*. Springer Singapore, 2021, p. 377–392. [Online]. Available: http://dx.doi.org/10.1007/978-981-16-0733-2_26

[3] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, p. 115–118, Jan. 2017. [Online]. Available: http://dx.doi.org/10.1038/nature21056

[4] T. Rahman, A. Khandakar, Y. Qiblawey, A. Tahir, S. Kiranyaz, S. B. Abul Kashem, M. T. Islam, S. Al Maadeed, S. M. Zughaier, M. S. Khan, and M. E. Chowdhury, "Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images," *Computers in Biology and Medicine*, vol. 132, p. 104319, May 2021. [Online]. Available: http://dx.doi.org/10.1016/j.compbiomed.2021.104319

[5] M. E. H. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. A. Emadi, M. B. I. Reaz, and M. T. Islam, "Can AI Help in Screening Viral and COVID-19 Pneumonia?" *IEEE Access*, vol. 8, p. 132665–132676, 2020. [Online]. Available: http://dx.doi.org/10.1109/ACCESS.2020.3010287

[6] H. Chan, L. M. Hadjiiski, and R. K. Samala, "Computer-aided diagnosis in the era

of deep learning," *Medical Physics*, vol. 47, no. 5, May 2020. [Online]. Available: http://dx.doi.org/10.1002/mp.13764

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[8] W. Yao, J. Bai, W. Liao, Y. Chen, M. Liu, and Y. Xie, "From CNN to Transformer: A Review of Medical Image Segmentation Models," *Journal of Imaging Informatics in Medicine*, vol. 37, no. 4, p. 1529–1547, Mar. 2024. [Online]. Available: http://dx.doi.org/10.1007/s10278-024-00981-7

[9] L. Gai, M. Xing, W. Chen, Y. Zhang, and X. Qiao, "Comparing CNN-based and transformer-based models for identifying lung cancer: which is more effective?" *Multimedia Tools and Applications*, vol. 83, no. 20, p. 59253–59269, Dec. 2023. [Online]. Available: http://dx.doi.org/10.1007/s11042-023-17644-4

[10] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: a literature review," *BMC Medical Imaging*, vol. 22, no. 1, Apr. 2022. [Online]. Available: http://dx.doi.org/10.1186/s12880-022-00793-7

[11] V. Cheplygina, M. de Bruijne, and J. P. Pluim, "Not-so-supervised: A survey of semi-supervised, multi-instance, and transfer learning in medical image analysis," *Medical Image Analysis*, vol. 54, pp. 280–296, 2019.

[12] X. Han, Z. Zhang, N. Ding, Y. Gu, X. Liu, Y. Huo, J. Qiu, Y. Yao, A. Zhang, L. Zhang, W. Han, M. Huang, Q. Jin, Y. Lan, Y. Liu, Z. Liu, Z. Lu, X. Qiu, R. Song, J. Tang, J.-R. Wen, J. Yuan, W. X. Zhao, and J. Zhu, "Pre-trained models: Past, present and future," *AI Open*, vol. 2, pp. 225–250, 2021.

[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[15] A. Ke, W. Ellsworth, O. Banerjee, A. Y. Ng, and P. Rajpurkar, "CheXtransfer: performance and parameter efficiency of ImageNet models for chest X-Ray interpretation," in *Proceedings of the Conference on Health, Inference, and Learning*, ser. ACM CHIL '21. ACM, Apr 2021.

[16] S. Woerner and C. F. Baumgartner, "Navigating Data Scarcity Using Foundation Models: A Benchmark of Few-Shot and Zero-Shot Learning Approaches in Medical Imaging," in *Foundation Models for General Medical AI*, Z. Deng, Y. Shen, H. J. Kim, W.-K. Jeong, A. I. Aviles-Rivero, J. He, and S. Zhang, Eds. Cham: Springer Nature Switzerland, 2025, pp. 30–39. [Online]. Available: http://dx.doi.org/10.1007/978-3-031-73471-7_4

[17] Y. Qiu, F. Lin, W. Chen, and M. Xu, "Pre-training in Medical Data: A Survey," *Machine Intelligence Research*, vol. 20, no. 2, pp. 147–179, 2023.

[18] S. Atasever, N. Azginoglu, D. S. Terzi, and R. Terzi, "A comprehensive survey of deep learning research on medical image analysis with focus on transfer learning," *Clinical Imaging*, vol. 94, p. 18–41, Feb 2023.

[19] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong, A. Tupini, Y. Wang, M. Mazzola, S. Shukla, L. Liden, J. Gao, M. P. Lungren, T. Naumann, S. Wang, and H. Poon, "BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs," 2023. [Online]. Available: https://arxiv.org/abs/2303.00915

[20] X. Mei, Z. Liu, P. M. Robson, B. Marinelli, M. Huang, A. Doshi, A. Jacobi, C. Cao, K. E. Link, T. Yang, Y. Wang, H. Greenspan, T. Deyer, Z. A. Fayad, and Y. Yang, "RadImageNet: An Open Radiologic Deep Learning Research Dataset for Effective Transfer Learning," *Radiology: Artificial Intelligence*, vol. 4, no. 5, Sep. 2022. [Online]. Available: http://dx.doi.org/10.1148/ryai.210315

[21] J. Silva-Rodríguez, J. Dolz, and I. B. Ayed, "Towards Foundation Models and Few-Shot Parameter-Efficient Fine-Tuning for Volumetric Organ Segmentation," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 Workshops*. Cham: Springer Nature Switzerland, 2023, pp. 213–224. [Online]. Available: http://dx.doi.org/10.1007/978-3-031-47401-9_21

[22] H.-Y. Zhou, S. Yu, C. Bian, Y. Hu, K. Ma, and Y. Zheng, "Comparing to Learn: Surpassing ImageNet Pretraining on Radiographs by Comparing Image Representations," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*. Cham: Springer International Publishing, 2020, pp. 398–407. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-59710-8_39

[23] L. Alzubaidi, J. Santamaría, M. Manoufali, B. Mohammed, M. A. Fadhel, J. Zhang, A. H. Al-Timemy, O. Al-Shamma, and Y. Duan, "MedNet: Pre-trained Convolutional Neural Network Model for the Medical Imaging Tasks," *CoRR*, vol. abs/2110.06512, 2021.

[24] X. Yang, A. Chen, N. PourNejatian, H. C. Shin, K. E. Smith, C. Parisien, C. Compas, C. Martin, A. B. Costa, M. G. Flores, Y. Zhang, T. Magoc, C. A. Harle, G. Lipori, D. A. Mitchell, W. R. Hogan, E. A. Shenkman, J. Bian, and Y. Wu, "A large language model for electronic health records," *npj Digital Medicine*, vol. 5, no. 1, p. 194, Dec 2022. [Online]. Available: https://doi.org/10.1038/s41746-022-00742-2

[25] J. Wasserthal, H.-C. Breit, M. T. Meyer, M. Pradella, D. Hinck, A. W. Sauter, T. Heye, D. T. Boll, J. Cyriac, S. Yang, M. Bach, and M. Segeroth, "TotalSegmentator: Robust Segmentation of 104 Anatomic Structures in CT Images," *Radiology: Artificial Intelligence*, vol. 5, no. 5, Sep. 2023. [Online]. Available: http://dx.doi.org/10.1148/ryai.230024

[26] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods*, vol. 18, no. 2, p. 203–211, Dec. 2020. [Online]. Available: http://dx.doi.org/10.1038/s41592-020-01008-z

[27] H. Wang, S. Guo, J. Ye, Z. Deng, J. Cheng, T. Li, J. Chen, Y. Su, Z. Huang, Y. Shen, B. Fu, S. Zhang, J. He, and Y. Qiao, "SAM-Med3D: Towards General-purpose Segmentation Models for Volumetric Medical Images," 2023. [Online]. Available: https://arxiv.org/abs/2310.15161

[28] Q. Xie, Q. Chen, A. Chen, C. Peng, Y. Hu, F. Lin, X. Peng, J. Huang, J. Zhang, V. Keloth, X. Zhou, L. Qian, H. He, D. Shung, L. Ohno-Machado, Y. Wu, H. Xu, and J. Bian, "Me LLaMA: Foundation Large Language Models for Medical Applications," 2024. [Online]. Available: https://arxiv.org/abs/2402.12749

[29] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *CoRR*, vol. abs/1810.04805, 2018.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2016.

[31] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jul. 2017.

[32] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 97. PMLR, 09–15 Jun 2019, pp. 6105–6114. [Online]. Available: https://proceedings.mlr.press/v97/tan19a.html

[33] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, "CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 590–597, Jul. 2019. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/3834

[34] K. Yan, X. Wang, L. Lu, and R. M. Summers, "DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning," *Journal of Medical Imaging*, vol. 5, no. 3, p. 036501, 2018. [Online]. Available: https://doi.org/10.1117/1.JMI.5.3.036501

[35] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[36] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Scientific Data*, vol. 6, no. 1, p. 317, Dec 2019. [Online]. Available: https://doi.org/10.1038/s41597-019-0322-0

[37] I. Siragusa, S. Contino, M. L. Ciura, R. Alicata, and R. Pirrone, "MedPix 2.0: A Comprehensive Multimodal Biomedical Data set for Advanced AI Applications with Retrieval Augmented Generation and Knowledge Graphs," 2025. [Online]. Available: https://arxiv.org/abs/2407.02994

[38] P. Rajpurkar, J. Irvin, A. Bagul, D. Ding, T. Duan, H. Mehta, B. Yang, K. Zhu, D. Laird, R. L. Ball, C. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, "MURA: Large Dataset for Abnormality Detection in Musculoskeletal Radiographs," 2018. [Online]. Available: https://arxiv.org/abs/1712.06957

[39] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner, "Open Access Series of Imaging Studies (OASIS): Cross-sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults," *Journal of Cognitive Neuroscience*, vol. 19, no. 9, pp. 1498–1507, 09 2007. [Online]. Available: https://doi.org/10.1162/jocn.2007.19.9.1498

[40] P. J. LaMontagne, S. Keefe, W. Lauren, C. Xiong, E. A. Grant, K. L. Moulder, J. C. Morris, T. L. Benzinger, and D. S. Marcus, "P3-083: OASIS-3: LONGITUDINAL NEUROIMAGING, CLINICAL, AND COGNITIVE

DATASET FOR NORMAL AGING AND ALZHEIMER'S DISEASE," *Alzheimer's and Dementia*, vol. 14, no. 7S_Part_20, Jul. 2018. [Online]. Available: http://dx.doi.org/10.1016/j.jalz.2018.06.1439

[41] M. Napravnik, F. Hržić, S. Tschauner, and I. Štajduhar, "Building RadiologyNET: an unsupervised approach to annotating a large-scale multimodal medical database," *BioData Mining*, vol. 17, no. 1, Jul. 2024. [Online]. Available: http://dx.doi.org/10.1186/s13040-024-00373-1

[42] DICOM Standards Committee, " DICOM Standard ," https://www.dicomstandard.org/, 2024, accessed on 5 Jul 2024.

[43] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning*, ser. ICML'20. JMLR.org, 2020.

[44] L.-Y. Guo, A.-H. Wu, Y.-x. Wang, L.-p. Zhang, H. Chai, and X.-F. Liang, "Deep learning-based ovarian cancer subtypes identification using multi-omics data," *BioData Mining*, vol. 13, no. 1, Aug. 2020. [Online]. Available: http://dx.doi.org/10.1186/s13040-020-00222-x

[45] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, no. 4, pp. 433–459, jun 2010.

[46] M. Napravnik, R. Baždarić, D. Miletić, F. Hržić, S. Tschauner, M. Mamula, and I. Štajduhar, "Using Autoencoders to Reduce Dimensionality of DICOM Metadata," in *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, 2022, pp. 1–6.

[47] W. A. Qader, M. M. Ameen, and B. I. Ahmed, "An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges," in *2019 International Engineering Conference (IEC)*. IEEE, Jun. 2019. [Online]. Available: http://dx.doi.org/10.1109/iec47844.2019.8950616

[48] T. Wiatowski and H. Bolcskei, "A Mathematical Theory of Deep Convolutional Neural Networks for Feature Extraction," *IEEE Transactions on Information*

*Theory*, vol. 64, no. 3, p. 1845–1866, Mar. 2018. [Online]. Available: http://dx.doi.org/10.1109/tit.2017.2776228

[49] C. K. Enders, *Applied missing data analysis*, 2nd ed. London, England: Guilford Press, Oct. 2022.

[50] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Jun. 2016. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2016.308

[51] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," 2014. [Online]. Available: https://arxiv.org/abs/1409.1556

[52] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for MobileNetV3," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2019. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2019.00140

[53] E. Nagy, M. Janisch, F. Hržić, E. Sorantin, and S. Tschauner, "A pediatric wrist trauma X-ray dataset (GRAZPEDWRI-DX) for machine learning," *Scientific Data*, vol. 9, no. 1, May 2022.

[54] M. Nickparvar, "Brain Tumor MRI Dataset," https://www.kaggle.com/dsv/2645886, 2021. [Online]. Available: https://www.kaggle.com/dsv/2645886

[55] S. S. Halabi, L. M. Prevedello, J. Kalpathy-Cramer, A. B. Mamonov, A. Bilbily, M. Cicero, I. Pan, L. A. Pereira, R. T. Sousa, N. Abdala, F. C. Kitamura, H. H. Thodberg, L. Chen, G. Shih, K. Andriole, M. D. Kohli, B. J. Erickson, and A. E. Flanders, "The RSNA Pediatric Bone Age Machine Learning Challenge," *Radiology*, vol. 290, no. 2, pp. 498–503, Feb 2019.

[56] S. G. Armato, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. A. Kazerooni, H. MacMahon, E. J. R. van Beek, D. Yankelevitz, A. M. Biancardi, P. H. Bland, M. S. Brown, R. M. Engelmann, G. E. Laderach, D. Max, R. C. Pais, D. P. Qing,

R. Y. Roberts, A. R. Smith, A. Starkey, P. Batra, P. Caligiuri, A. Farooqi, G. W. Gladish, C. M. Jude, R. F. Munden, I. Petkovska, L. E. Quint, L. H. Schwartz, B. Sundaram, L. E. Dodd, C. Fenimore, D. Gur, N. Petrick, J. Freymann, J. Kirby, B. Hughes, A. Vande Casteele, S. Gupte, M. Sallam, M. D. Heath, M. H. Kuhn, E. Dharaiya, R. Burns, D. S. Fryd, M. Salganicoff, V. Anand, U. Shreter, S. Vastagh, B. Y. Croft, and L. P. Clarke, "The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans," *Medical Physics*, vol. 38, no. 2, pp. 915–931, Jan 2011.

[57] ——, "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge," *Medical Image Analysis*, vol. 42, pp. 1–13, Dec 2017.

[58] D. Miller and the pydicom contributors, "pydicom," 2024, python library for DICOM file parsing and manipulation. [Online]. Available: https://github.com/pydicom/pydicom

[59] Matthew Rocklin, "Dask: Parallel Computation with Blocked algorithms and Task Scheduling," in *Proceedings of the 14th Python in Science Conference*, 2015, pp. 126 – 132. [Online]. Available: https://doi.org/10.25080/Majora-7b98e3ed-013

[60] A. Shetty, C. Hacking, and A. Campos, "Positron emission tomography," Jun. 2014. [Online]. Available: http://dx.doi.org/10.53347/rID-29716

[61] K. Bhaskaran and L. Smeeth, "What is the difference between missing completely at random and missing at random?" *International Journal of Epidemiology*, vol. 43, no. 4, pp. 1336–1339, apr 2014.

[62] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, "A survey on missing data in machine learning," *Journal of Big Data*, vol. 8, no. 1, oct 2021.

[63] P. Mildenberger, M. Eichelberg, and E. Martin, "Introduction to the DICOM standard," *European Radiology*, vol. 12, no. 4, pp. 920–927, sep 2001.

[64] D. J. Stekhoven and P. Bühlmann, "MissForest—non-parametric missing value imputation for mixed-type data," *Bioinformatics*, vol. 28, no. 1, pp. 112–118, oct 2011.

[65] F. Tang and H. Ishwaran, "Random forest missing data algorithms," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 10, no. 6, pp. 363–377, jun 2017.

[66] F. Hržić, M. Napravnik, R. Baždarić, I. Štajduhar, M. Mamula, D. Miletić, and S. Tschauner, "Estimation of Missing Parameters for DICOM to 8-bit X-ray Image Export," in *2022 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, 2022, pp. 1–6.

[67] M. Napravnik, N. Bakotić, F. Hržić, D. Miletić, and I. Štajduhar, "Estimation of Missing DICOM Windowing Parameters in High-Dynamic-Range Radiographs Using Deep Learning," *Mathematics*, vol. 13, no. 10, p. 1596, May 2025. [Online]. Available: http://dx.doi.org/10.3390/math13101596

[68] S. K. Thompson, C. E. Willis, K. T. Krugh, S. J. Shepard, and K. W. McEnery, "Implementing the DICOM Grayscale Standard Display Function for Mixed Hard- and Soft-Copy Operations," *Journal of Digital Imaging*, vol. 15, no. 0, pp. 27–32, apr 2002.

[69] N. Ljubešić, D. Boras, and O. Kubelka, "Retrieving information in Croatian: Building a simple and efficient rule-based stemmer," in *The Future of Information Sciences (INFuture 2007) : Digital information and heritage*. Odsjek za informacijske znanosti, Filozofski fakultet, Zagreb, 2007, pp. 313–320.

[70] A. F. Agarap, "Deep Learning using Rectified Linear Units (ReLU)," 2018.

[71] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," 2014. [Online]. Available: https://arxiv.org/abs/1412.6980

[72] N. Tajbakhsh, L. Jeyaseelan, Q. Li, J. N. Chiang, Z. Wu, and X. Ding, "Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation," *Medical Image Analysis*, vol. 63, p. 101693, jul 2020.

[73] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Aafssisted Intervention – MICCAI 2015*. Springer International Publishing, 2015, pp. 234–241.

[74] M. Z. Alom, M. Hasan, C. Yakopcic, T. M. Taha, and V. K. Asari, "Recurrent Residual Convolutional Neural Network based on U-Net (R2U-Net) for Medical Image Segmentation," 2018.

[75] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, "Attention U-Net: Learning Where to Look for the Pancreas," 2018. [Online]. Available: https://arxiv.org/abs/1804.03999

[76] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient Estimation of Word Representations in Vector Space, publisher = arXiv," 2013.

[77] N. C. Dang, M. N. Moreno-García, and F. De la Prieta, "Sentiment Analysis Based on Deep Learning: A Comparative Study," *Electronics*, vol. 9, no. 3, 2020.

[78] Q. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," in *Proceedings of the 31st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, E. P. Xing and T. Jebara, Eds., vol. 32(2). Bejing, China: PMLR, 22–24 Jun 2014, pp. 1188–1196.

[79] F. K. Khattak, S. Jeblee, C. Pou-Prom, M. Abdalla, C. Meaney, and F. Rudzicz, "A survey of word embeddings for clinical text," *Journal of Biomedical Informatics*, vol. 100, p. 100057, 2019.

[80] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text Classification Algorithms: A Survey," *Information*, vol. 10, no. 4, 2019.

[81] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P. S. Yu, and L. He, "A Survey on Text Classification: From Traditional to Deep Learning," *ACM Trans. Intell. Syst. Technol.*, vol. 13, no. 2, apr 2022.

[82] D. N. Nicholson, F. Alquaddoomi, V. Rubinetti, and C. S. Greene, "Changing word meanings in biomedical literature reveal pandemics and new technologies," *BioData Mining*, vol. 16, no. 1, May 2023.

[83] E. Anderson, Z. Bai, C. Bischof, L. S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, and D. Sorensen, "LAPACK Users' Guide," 1999. [Online]. Available: https://epubs.siam.org/doi/abs/10.1137/1.9780898719604

[84] R. B. Lehoucq, D. C. Sorensen, and C. Yang, *ARPACK users' guide: solution of large-scale eigenvalue problems with implicitly restarted Arnoldi methods*, 1998.

[85] P.-G. Martinsson, V. Rokhlin, and M. Tygert, "A randomized algorithm for the decomposition of matrices," *Applied and Computational Harmonic Analysis*, vol. 30, no. 1, pp. 47–68, jan 2011.

[86] C. Elkan, "Using the triangle inequality to accelerate k-means," in *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ser. ICML'03. AAAI Press, 2003, p. 147–153.

[87] L. Rdusseeun and P. Kaufman, "Clustering by means of medoids," in *Proceedings of the statistical data analysis based on the L1 norm conference, neuchatel, switzerland*, vol. 31, 1987, pp. 405–416.

[88] T. Kvålseth, "On Normalized Mutual Information: Measure Derivations and Properties," *Entropy*, vol. 19, no. 11, p. 631, nov 2017.

[89] R. L. Thorndike, "Who belongs in the family?" *Psychometrika*, vol. 18, no. 4, pp. 267–276, dec 1953.

[90] V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan, "Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior," in *2011 31st International Conference on Distributed Computing Systems Workshops*. IEEE, 2011, pp. 166–171.

[91] M. Wang, C. Lee, Z. Wei, H. Ji, Y. Yang, and C. Yang, "Clinical assistant decision-making model of tuberculosis based on electronic health records," *BioData Mining*, vol. 16, no. 1, Mar. 2023.

[92] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[93] M. Mikulić, D. Vičević, E. Nagy, M. Napravnik, I. Štajduhar, S. Tschauner, and F. Hržić, "Balancing Performance and Interpretability in Medical Image Analysis: Case study of Osteopenia," *Journal of Imaging Informatics in Medicine*, Jul. 2024. [Online]. Available: http://dx.doi.org/10.1007/s10278-024-01194-8

[94] D. Müller, I. Soto-Rey, and F. Kramer, "Towards a guideline for evaluation metrics in medical image segmentation," *BMC Research Notes*, vol. 15, no. 1, Jun. 2022. [Online]. Available: http://dx.doi.org/10.1186/s13104-022-06096-y

[95] B. B. Schultz, "Levene's Test for Relative Variation," *Systematic Biology*, vol. 34, no. 4, pp. 449–456, 12 1985. [Online]. Available: https://doi.org/10.1093/sysbio/34.4.449

[96] W. H. Kruskal and W. A. Wallis, "Use of Ranks in One-Criterion Variance Analysis," *Journal of the American Statistical Association*, vol. 47, no. 260, pp. 583–621, 1952. [Online]. Available: https://www.tandfonline.com/doi/abs/10.1080/01621459.1952.10483441

[97] B. Courty, V. Schmidt, S. Luccioni, Goyal-Kamal, MarionCoutarel, B. Feld, J. Lecourt, LiamConnell, A. Saboni, Inimaz, supatomic, M. Léval, L. Blanche, A. Cruveiller, ouminasara, F. Zhao, A. Joshi, A. Bogroff, H. de Lavoreille, N. Laskaris, E. Abati, D. Blank, Z. Wang, A. Catovic, M. Alencon, M. Stęchły, C. Bauer, L. O. N. de Araújo, JPW, and MinervaBooks, "mlco2/codecarbon: v2.4.1," May 2024. [Online]. Available: https://doi.org/10.5281/zenodo.11171501

[98] M. Benchari and M. W. Totaro, *MRI Brain Cancer Image Detection: Application of*

*an Integrated U-Net and ResNet50 Architecture*. Springer Nature Switzerland, 2024, p. 104–108. [Online]. Available: http://dx.doi.org/10.1007/978-3-031-66535-6_12

[99] W. Liu, J. Luo, Y. Yang, W. Wang, J. Deng, and L. Yu, "Automatic lung segmentation in chest X-ray images using improved U-Net," *Scientific Reports*, vol. 12, no. 1, May 2022. [Online]. Available: http://dx.doi.org/10.1038/s41598-022-12743-y

[100] E. Stevens and L. Antiga, *Deep learning with PyTorch*. New York, NY: Manning Publications, Oct 2020.

[101] K. He, R. Girshick, and P. Dollar, "Rethinking ImageNet Pre-Training," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Oct 2019.

[102] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal, M. Schaekermann, A. Wang, M. Amin, S. Lachgar, P. Mansfield, S. Prakash, B. Green, E. Dominowska, B. A. y Arcas, N. Tomasev, Y. Liu, R. Wong, C. Semturs, S. S. Mahdavi, J. Barral, D. Webster, G. S. Corrado, Y. Matias, S. Azizi, A. Karthikesalingam, and V. Natarajan, "Towards Expert-Level Medical Question Answering with Large Language Models," 2023. [Online]. Available: https://arxiv.org/abs/2305.09617

[103] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *International Journal of Computer Vision*, vol. 128, no. 2, p. 336–359, Oct. 2019. [Online]. Available: http://dx.doi.org/10.1007/s11263-019-01228-7

[104] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, "The false hope of current approaches to explainable artificial intelligence in health care," *The Lancet Digital Health*, vol. 3, no. 11, p. e745–e750, Nov. 2021. [Online]. Available: http://dx.doi.org/10.1016/S2589-7500(21)00208-9

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF CODE LISTINGS

# LIST OF ABBREVIATIONS

| | |
|---|---|
| AE | Autoencoder |
| ANOVA | Analysis of Variance |
| BPE | Body Part Examined |
| BTMR | Brain Tumor MRI |
| BoW | Bag of Words |
| CAE | Convolutional Autoencoder |
| CLIP | Contrastive Language–Image Pretraining |
| CNN | Convolutional Neural Network |
| CR | Computed Radiography |
| CSV | Comma-Separated Values |
| CT | Computed Tomography |
| DICOM | Digital Imaging and Communications in Medicine |
| Grad-CAM | Gradient-weighted Class Activation Mapping |
| HS | Homogeneity Score |
| IoU | Intersection over Union |
| LUNA | LUng Nodule Analysis |
| MAE | Mean Absolute Error |
| MR | Magnetic Resonance |
| MSE | Mean Squared Error |
| MWU | Mann–Whitney U test |
| NM | Nuclear Medicine |
| NMI | Normalised Mutual Information |
| PACS | Picture Archiving and Communication System |
| PBA | Pediatric Bone Age |
| PCA | Principal Component Analysis |
| PNG | Portable Network Graphics |
| PV-DBOW | Paragraph Vector - Distributed Bag of Words |
| PV-DM | Paragraph Vector - Distributed Memory |
| RF | Radiofluoroscopy |
| RMSE | Root Mean Squared Error |
| ReLU | Rectified Linear Unit |
| TF-IDF | Term Frequency-Inverse Document Frequency |
| TL | Transfer Learning |
| XA | X-ray Angiography |

# APPENDIX

# A. Performance on the validation set of downstream tasks

**Table A.1:** RSNA PBA: results on the validation set, by learning rate. The shown metric is Mean Absolute Error, averaged across five runs.

| Challenge | Model | Learning Rate | | | | Avg. |
|---|---|---|---|---|---|---|
| | | $10^{-2}$ | $10^{-3}$ | $10^{-4}$ | $10^{-5}$ | |
| PBA EfficientNetB3 (avg.) | ImageNet | 17.04 ± 9.9 | 9.88 ± 0.5 | 12.67 ± 0.5 | 13.96 ± 0.2 | 13.39 ± 2.6 |
| | RadiologyNET | 12.19 ± 1.9 | 9.46 ± 0.1 | 10.53 ± 0.1 | 12.61 ± 0.0 | 11.2 ± 1.3 |
| | Baseline | 34.35 ± 9.4 | 21.91 ± 10.5 | 14.42 ± 2.0 | 23.97 ± 4.1 | 23.66 ± 7.1 |
| PBA EfficientNetB3 (best) | ImageNet | 9.995 | 9.041 | 11.808 | 13.794 | 11.2 ± 1.8 |
| | RadiologyNET | 10.379 | 9.27 | 10.426 | 12.589 | 10.7 ± 1.2 |
| | Baseline | 25.411 | 9.882 | 12.435 | 20.901 | 17.2 ± 6.3 |
| PBA InceptionV3 (avg.) | ImageNet | 278.87 ± 469.2 | 10.46 ± 0.6 | 10.46 ± 1.3 | 11.97 ± 0.5 | 77.94 ± 116.0 |
| | RadiologyNET | 10.16 ± 0.2 | 10.54 ± 0.2 | 12.83 ± 0.2 | 36.45 ± 0.0 | 17.5 ± 11.0 |
| | Baseline | 11.12 ± 0.8 | 9.83 ± 0.2 | 12.45 ± 1.4 | 20.0 ± 2.4 | 13.35 ± 3.9 |
| PBA InceptionV3 (best) | ImageNet | 36.553 | 9.747 | 9.444 | 11.268 | 16.8 ± 11.5 |
| | RadiologyNET | 9.861 | 10.381 | 12.627 | 36.433 | 17.3 ± 11.1 |
| | Baseline | 10.171 | 9.583 | 10.719 | 17.295 | 11.9 ± 3.1 |

**Table A.2:** GRAZPEDWRI-DX: results on the validation set, by learning rate. The shown metric is F1-score, averaged across five runs.

| Challenge | Model | Learning Rate | | | | Avg. |
|---|---|---|---|---|---|---|
| | | $10^{-2}$ | $10^{-3}$ | $10^{-4}$ | $10^{-5}$ | |
| GRAZPEDWRI-DX DenseNet121 (avg.) | ImageNet | 91.4 ± 0.2 | 93.1 ± 0.5 | 92.8 ± 0.6 | 90.8 ± 1.1 | 92.0 ± 1.0 |
| | RadiologyNET | 90.8 ± 0.8 | 91.8 ± 0.4 | 92.0 ± 1.0 | 92.4 ± 0.3 | 91.7 ± 0.6 |
| | Baseline | 89.1 ± 1.4 | 90.3 ± 2.1 | 89.4 ± 0.6 | 86.9 ± 0.8 | 88.9 ± 1.3 |
| GRAZPEDWRI-DX DenseNet121 (best) | ImageNet | 91.6 | 93.5 | 93.5 | 92.1 | 92.7 ± 0.8 |
| | RadiologyNET | 91.7 | 92.2 | 93.0 | 92.6 | 92.4 ± 0.5 |
| | Baseline | 90.8 | 92.6 | 90.0 | 88.2 | 90.4 ± 1.6 |
| GRAZPEDWRI-DX ResNet34 (avg.) | ImageNet | 89.6 ± 0.4 | 92.6 ± 1.2 | 92.9 ± 0.6 | 91.0 ± 1.7 | 91.5 ± 1.3 |
| | RadiologyNET | 86.9 ± 3.8 | 91.2 ± 0.9 | 90.5 ± 1.2 | 90.3 ± 0.5 | 89.7 ± 1.6 |
| | Baseline | 79.1 ± 13.3 | 88.4 ± 0.8 | 88.8 ± 0.6 | 84.8 ± 2.2 | 85.3 ± 3.9 |
| GRAZPEDWRI-DX ResNet34 (best) | ImageNet | 90.1 | 94.3 | 93.5 | 93.0 | 92.7 ± 1.6 |
| | RadiologyNET | 90.1 | 92.4 | 91.4 | 90.6 | 91.1 ± 0.9 |
| | Baseline | 90.1 | 89.3 | 89.3 | 87.9 | 89.2 ± 0.8 |

**Table A.3:** COVID-19: results on the validation set, by learning rate. The shown metric is F1-score, averaged across five runs.

| Challenge | Model | Learning Rate | | | | Avg. |
|---|---|---|---|---|---|---|
| | | $10^{-2}$ | $10^{-3}$ | $10^{-4}$ | $10^{-5}$ | |
| COVID-19 | ImageNet | 91.3 ± 3.5 | 98.0 ± 1.0 | 98.3 ± 0.4 | 83.6 ± 0.2 | 92.8 ± 6.0 |
| MobileNetV3Large | RadiologyNET | 94.2 ± 1.1 | 97.9 ± 0.8 | 98.3 ± 0.4 | 97.7 ± 0.1 | 97.0 ± 1.6 |
| (avg.) | Baseline | 83.7 ± 4.2 | 95.0 ± 1.1 | 95.5 ± 1.9 | 93.2 ± 1.2 | 91.8 ± 4.8 |
| COVID-19 | ImageNet | 94.7 | 99.0 | 98.7 | 83.8 | 94.0 ± 6.2 |
| MobileNetV3Large | RadiologyNET | 96.1 | 98.7 | 98.7 | 97.8 | 97.8 ± 1.1 |
| (best) | Baseline | 91.1 | 96.7 | 97.5 | 94.2 | 94.9 ± 2.5 |
| COVID-19 | ImageNet | 93.3 ± 5.5 | 99.0 ± 0.3 | 99.3 ± 0.2 | 98.6 ± 0.0 | 97.6 ± 2.5 |
| ResNet18 | RadiologyNET | 97.4 ± 0.4 | 98.3 ± 0.2 | 98.4 ± 0.2 | 98.3 ± 0.1 | 98.1 ± 0.4 |
| (avg.) | Baseline | 95.1 ± 1.4 | 97.4 ± 0.5 | 96.5 ± 0.6 | 97.0 ± 0.6 | 96.5 ± 0.9 |
| COVID-19 | ImageNet | 98.0 | 99.4 | 99.6 | 98.6 | 98.9 ± 0.6 |
| ResNet18 | RadiologyNET | 97.9 | 98.5 | 98.6 | 98.3 | 98.3 ± 0.3 |
| (best) | Baseline | 97.2 | 98.0 | 97.3 | 97.7 | 97.6 ± 0.3 |

**Table A.4:** Brain Tumor MRI: results on the validation set, by learning rate. The shown metric is F1-score, averaged across five runs.

| Challenge | Model | Learning Rate | | | | Avg. |
|---|---|---|---|---|---|---|
| | | $10^{-2}$ | $10^{-3}$ | $10^{-4}$ | $10^{-5}$ | |
| BTMR | ImageNet | 95.1 ± 1.3 | 99.2 ± 0.5 | 99.6 ± 0.2 | 99.7 ± 0.1 | 98.4 ± 1.9 |
| ResNet50 | RadiologyNET | 96.0 ± 0.7 | 98.8 ± 0.5 | 99.6 ± 0.2 | 99.5 ± 0.2 | 98.5 ± 1.5 |
| (avg.) | Baseline | 96.5 ± 2.3 | 97.8 ± 1.1 | 98.8 ± 0.5 | 95.6 ± 1.8 | 97.2 ± 1.2 |
| BTMR | ImageNet | 96.4 | 99.6 | 99.9 | 99.9 | 98.9 ± 1.5 |
| ResNet50 | RadiologyNET | 96.9 | 99.2 | 99.8 | 99.6 | 98.9 ± 1.2 |
| (best) | Baseline | 98.3 | 98.8 | 99.4 | 97.4 | 98.5 ± 0.7 |
| BTMR | ImageNet | 12.6 ± 6.7 | 97.3 ± 0.6 | 99.3 ± 0.2 | 99.0 ± 0.4 | 77.0 ± 37.2 |
| MobileNetV3Small | RadiologyNET | 20.2 ± 9.8 | 94.8 ± 1.4 | 99.3 ± 0.2 | 99.0 ± 0.2 | 78.3 ± 33.6 |
| (avg.) | Baseline | 17.2 ± 10.5 | 87.4 ± 8.7 | 96.9 ± 1.6 | 94.4 ± 1.8 | 74.0 ± 33.0 |
| BTMR | ImageNet | 24.6 | 98.2 | 99.5 | 99.2 | 80.4 ± 32.2 |
| MobileNetV3Small | RadiologyNET | 29.2 | 96.4 | 99.6 | 99.2 | 81.1 ± 30.0 |
| (best) | Baseline | 29.6 | 92.6 | 98.9 | 96.5 | 79.4 ± 28.8 |

**Table A.5:** Results and $p$-values of statistical tests. PBA was compared using MAE, LUNA using Dice score, and other challenges were compared using F1-score. Values with significant differences ($p < 0.05$) are emphasised.

| Challenge | Kruskal-Wallis (p-value) | Mann-Whitney U test | MWU (p-value) |
|---|---|---|---|
| LUNA U-Net | **0.009** | ImageNet vs RadiologyNET | 1 |
| | | ImageNet vs Baseline | **0.024** |
| | | RadiologyNET vs Baseline | **0.024** |
| LUNA U-Net-EfficientNetB4 | 0.561 | ImageNet vs RadiologyNET | - |
| | | ImageNet vs Baseline | - |
| | | RadiologyNET vs Baseline | - |
| LUNA U-Net-ResNet50 | **0.011** | ImageNet vs RadiologyNET | 0.667 |
| | | ImageNet vs Baseline | 0.095 |
| | | RadiologyNET vs Baseline | **0.024** |
| LUNA U-Net-VGG16 | 0.108 | ImageNet vs RadiologyNET | - |
| | | ImageNet vs Baseline | - |
| | | RadiologyNET vs Baseline | - |
| PBA EfficientNetB3 | **0.014** | ImageNet vs RadiologyNET | 0.667 |
| | | ImageNet vs Baseline | 0.167 |
| | | RadiologyNET vs Baseline | **0.024** |
| PBA InceptionV3 | 0.185 | ImageNet vs RadiologyNET | - |
| | | ImageNet vs Baseline | - |
| | | RadiologyNET vs Baseline | - |
| GRAZPEDWRI-DX DenseNet121 | 0.063 | ImageNet vs RadiologyNET | - |
| | | ImageNet vs Baseline | - |
| | | RadiologyNET vs Baseline | - |
| GRAZPEDWRI-DX ResNet34 | **0.008** | ImageNet vs RadiologyNET | 0.286 |
| | | ImageNet vs Baseline | **0.024** |
| | | RadiologyNET vs Baseline | 0.095 |
| BTMR MobileNetV3Small | **0.017** | ImageNet vs RadiologyNET | 1 |
| | | ImageNet vs Baseline | **0.048** |
| | | RadiologyNET vs Baseline | **0.048** |
| BTMR ResNet50 | **0.008** | ImageNet vs RadiologyNET | 1 |
| | | ImageNet vs Baseline | **0.024** |
| | | RadiologyNET vs Baseline | **0.036** |
| COVID-19 MobileNetV3Large | **0.009** | ImageNet vs RadiologyNET | 1 |
| | | ImageNet vs Baseline | **0.047** |
| | | RadiologyNET vs Baseline | **0.024** |
| COVID-19 ResNet18 | **0.008** | ImageNet vs RadiologyNET | 1 |
| | | ImageNet vs Baseline | **0.035** |
| | | RadiologyNET vs Baseline | **0.035** |

**Table A.6:** Results and $p$-values of statistical tests when comparing the total epoch count until convergence. Values with significant differences ($p < 0.05$) are emphasised.

| **Challenge** | **Kruskal-Wallis (p-value)** | **Mann-Whitney U test** | **MWU (p-value)** |
|---|---|---|---|
| PBA EfficientNetB3 | **0.011** | ImageNet VS RadiologyNET | **0.033** |
| | | ImageNet VS Baseline | 0.667 |
| | | RadiologyNET vs Baseline | 0.122 |
| PBA InceptionV3 | **0.039** | ImageNet VS RadiologyNET | 0.139 |
| | | ImageNet VS Baseline | 0.102 |
| | | RadiologyNET vs Baseline | 1 |
| GRAZPEDWRI-DX DenseNet121 | **0.005** | ImageNet VS RadiologyNET | 0.070 |
| | | ImageNet VS Baseline | 0.225 |
| | | RadiologyNET vs Baseline | **0.033** |
| GRAZPEDWRI-DX ResNet34 | 0.199 | ImageNet VS RadiologyNET | - |
| | | ImageNet VS Baseline | **-** |
| | | RadiologyNET vs Baseline | - |
| BTMR MobileNetV3Small | 0.063 | ImageNet VS RadiologyNET | - |
| | | ImageNet VS Baseline | **-** |
| | | RadiologyNET vs Baseline | **-** |
| BTMR ResNet50 | **0.011** | ImageNet VS RadiologyNET | 0.052 |
| | | ImageNet VS Baseline | 1 |
| | | RadiologyNET vs Baseline | **0.028** |
| COVID-19 MobileNetV3Large | 0.326 | ImageNet VS RadiologyNET | - |
| | | ImageNet VS Baseline | **-** |
| | | RadiologyNET vs Baseline | **-** |
| COVID-19 ResNet18 | **0.002** | ImageNet VS RadiologyNET | **0.020** |
| | | ImageNet VS Baseline | **0.022** |
| | | RadiologyNET vs Baseline | 0.103 |

# B. Grad-CAM heatmaps

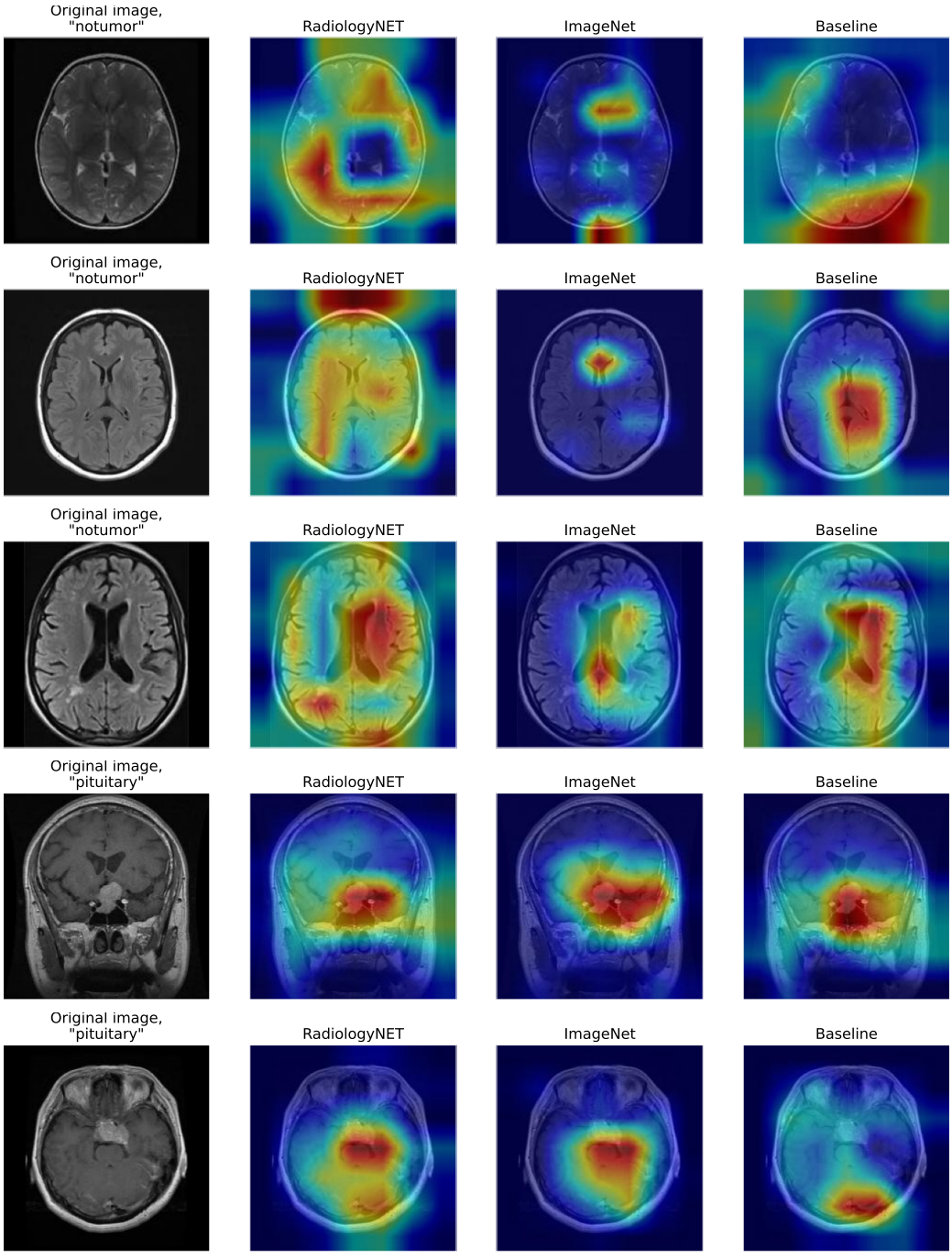



**Figure B.1:** Brain Tumor MRI ResNet50: Grad-CAM heatmap examples of randomly selected images from the Brain Tumor MRI dataset.

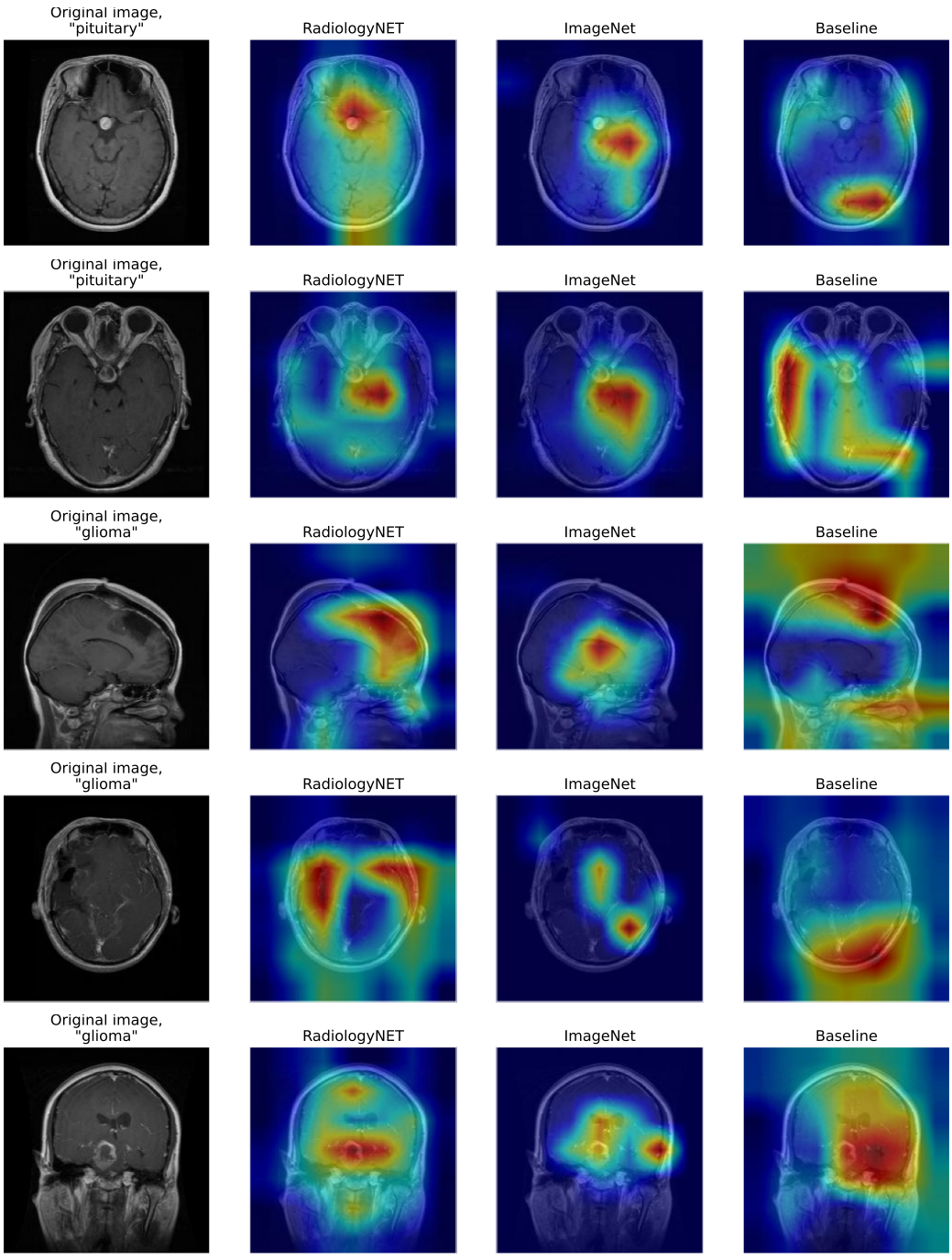



**Figure B.2:** Brain Tumor MRI ResNet50: Grad-CAM heatmap examples of randomly selected images from the Brain Tumor MRI dataset.

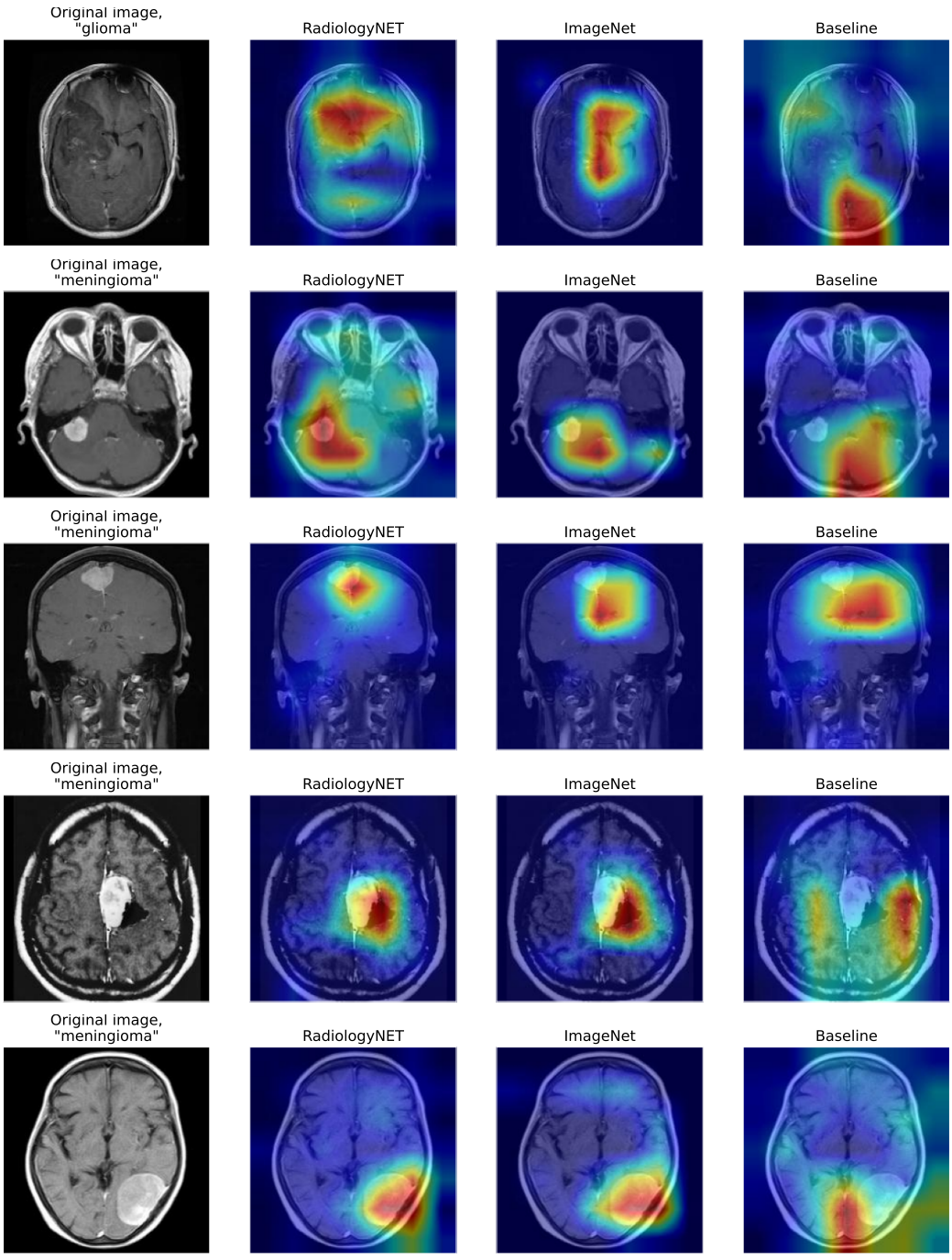



**Figure B.3:** Brain Tumor MRI ResNet50: Grad-CAM heatmap examples of randomly selected images from the Brain Tumor MRI dataset.

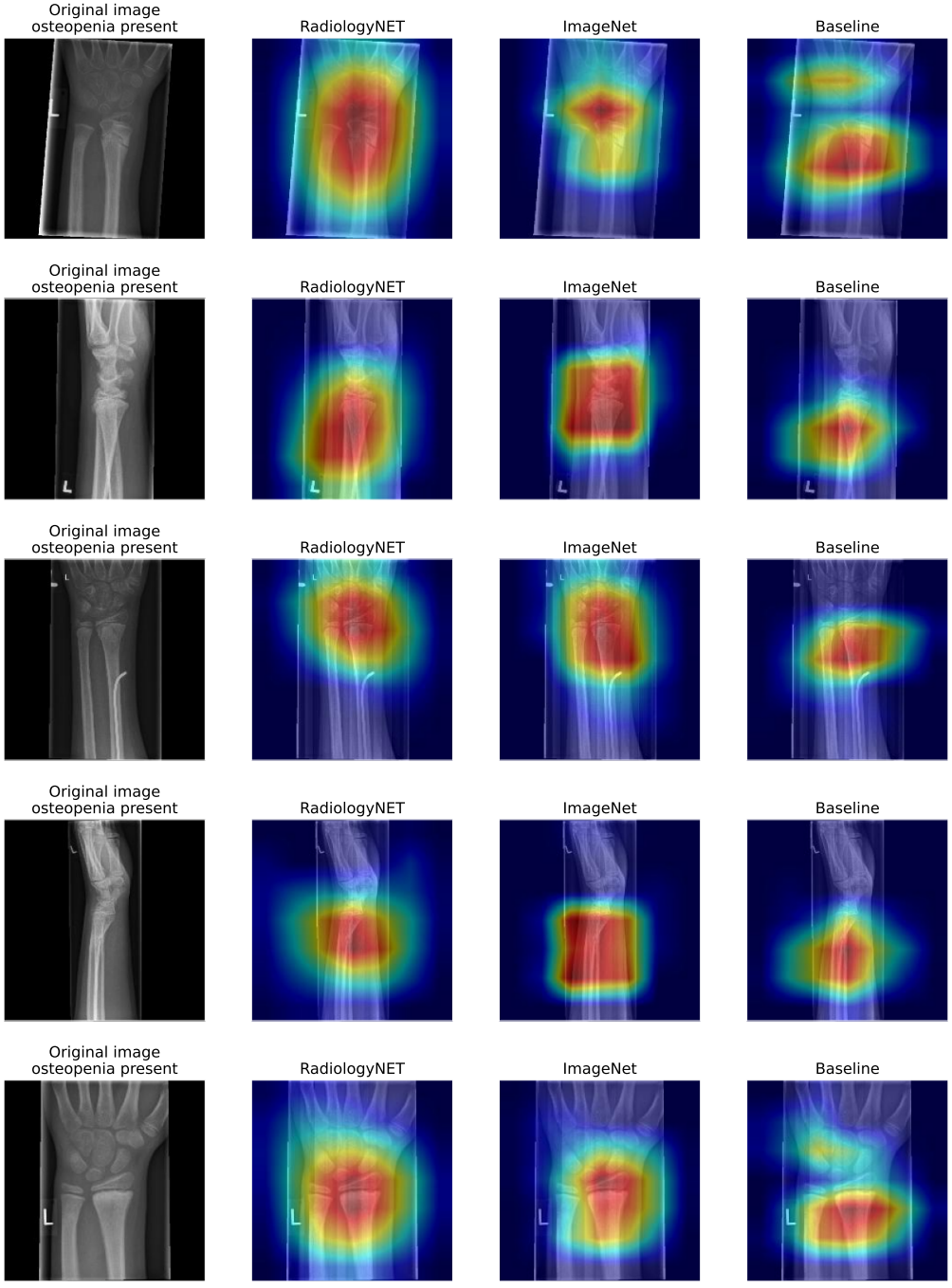



**Figure B.4:** GRAZPEDWRI-DX DenseNet121: Grad-CAM heatmap examples of randomly selected images from the GRAZPEDWRI-DX dataset.

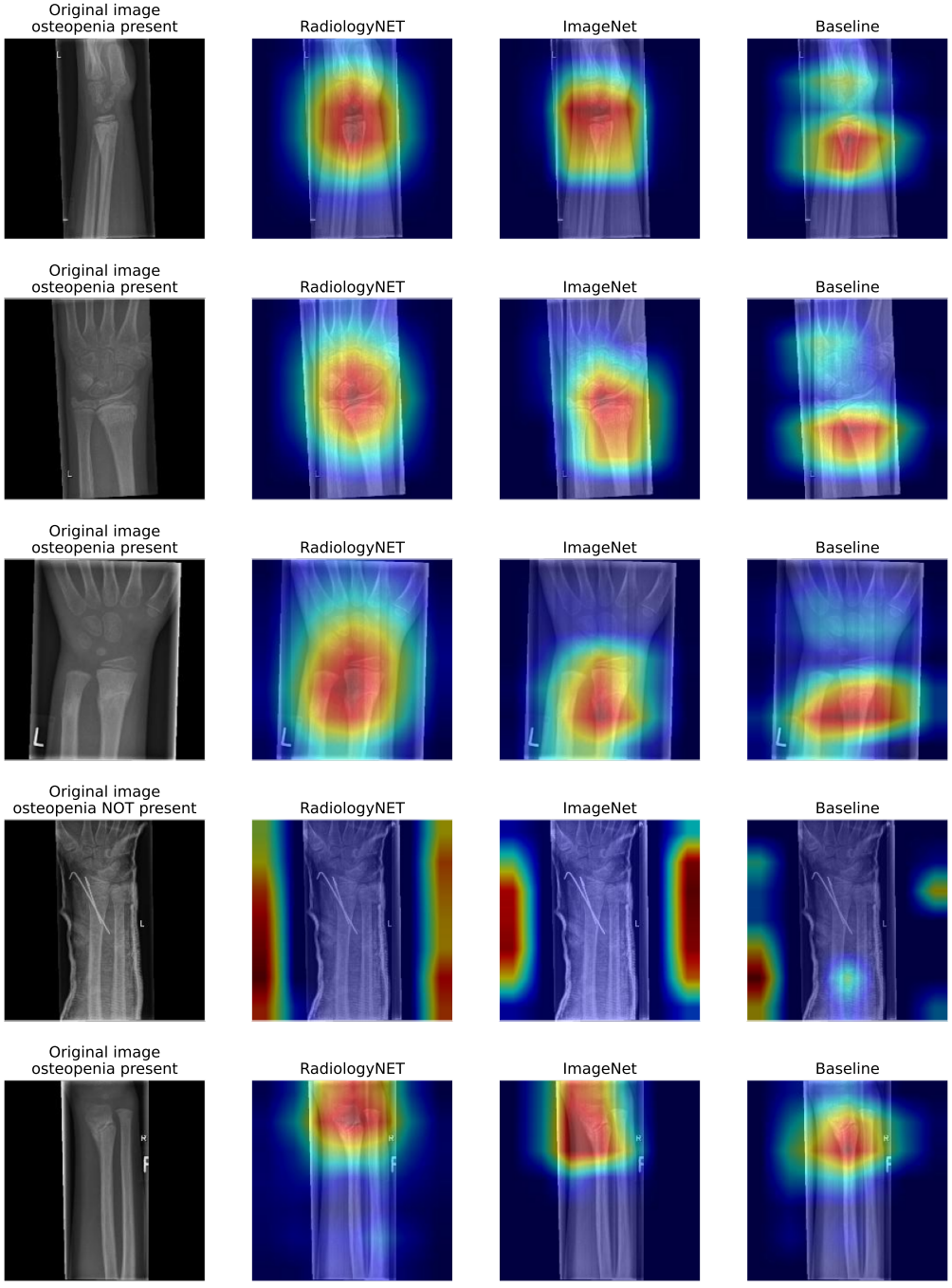



**Figure B.5:** GRAZPEDWRI-DX DenseNet121: Grad-CAM heatmap examples of randomly selected images from the GRAZPEDWRI-DX dataset.

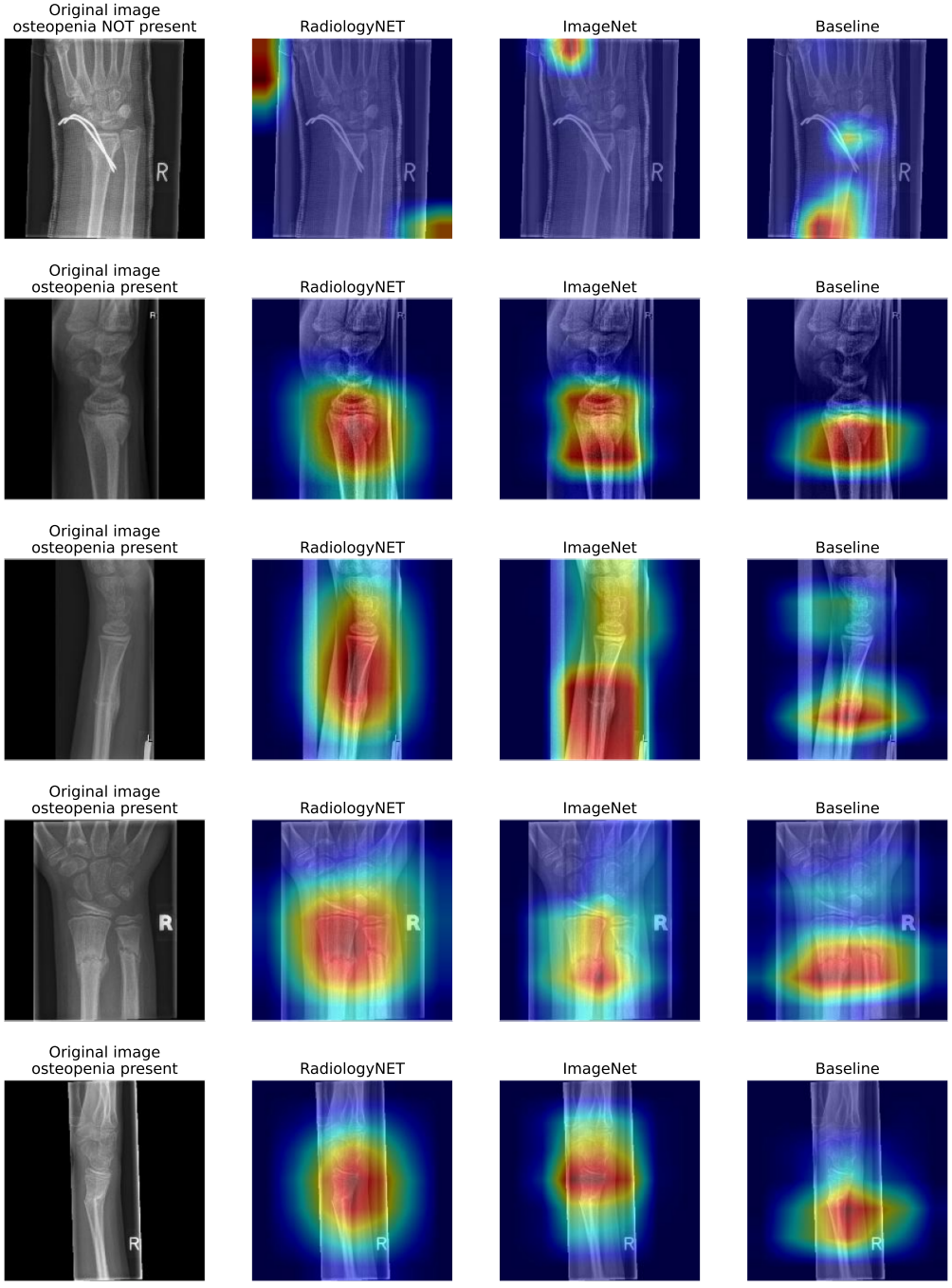



**Figure B.6:** GRAZPEDWRI-DX DenseNet121: Grad-CAM heatmap examples of randomly selected images from the GRAZPEDWRI-DX dataset.

# CURRICULUM VITAE

Mateja Napravnik was born on June 26$^{\text{th}}$, 1997, in Rijeka, Croatia. Having grown up in the town of Buje (Città di Buie), Istria, Croatia, in 2016 she enrolled in undergraduate studies at the University of Rijeka, Faculty of Engineering. Mateja earned her bachelor's degree in 2019 and her master's degree in 2021. Nearing the end of her master's degree, she received the Rector's Award for academic excellence and was named the Best Student at the University of Rijeka. That same year, she began her postgraduate doctoral studies in engineering sciences, specialising in computer science, at the Faculty of Engineering.

During her postgraduate doctoral studies, Mateja was employed as an assistant and a researcher at the University of Rijeka, Faculty of Engineering, Department of Computer Engineering. She actively participated in the project "Machine learning for knowledge transfer in medical radiology" under the leadership of Prof. Ivan Štajduhar, PhD. Her doctoral research focused on machine learning in medical image analysis, but she also frequently worked with students in various (extra)curricular activities, ranging from machine learning to web development.

# LIST OF PUBLICATIONS

[1] M. Napravnik, F. Hržić, S. Tschauner, and I. Štajduhar, 'Building RadiologyNET: an unsupervised approach to annotating a large-scale multimodal medical database', *BioData Mining*, vol. 17, no. 1, p. 22, 2024.

[2] M. Napravnik, N. Bakotić, F. Hržić, D. Miletić, and I. Štajduhar, 'Estimation of missing DICOM windowing parameters in high-dynamic-range radiographs using deep learning', *Mathematics*, vol. 13, no. 10, p. 1596, 2025.

[3] M. Napravnik, R. Baždarić, D. Miletić, F. Hržić, S. Tschauner, M. Mamula, and I. Štajduhar, 'Using Autoencoders to Reduce Dimensionality of DICOM Metadata', in 2022 *International Conference on Electrical, Computer, Communications and Mechatronics Engineering* (ICECCME), Maldives, 2022.

[4] A. Skoki, M. Napravnik, M. Polonijo, I. Štajduhar, and J. Lerga, 'Revolutionizing soccer injury management: Predicting muscle injury recovery time using ML', *Applied Sciences*, vol. 13, no. 10, p. 6222, 2023.

[5] M. Mikulić, D. Vičević, E. Nagy, M. Napravnik, I. Štajduhar, S. Tschauner, and F. Hržić, 'Balancing performance and interpretability in medical image analysis: Case study of osteopenia', *Journal of Imaging Informatics in Medicine*, vol. 38, no. 1, pp. 177–190, 2024.

[6] F. Hržić, M. Napravnik, R. Baždarić, I. Štajduhar, M. Mamula, D. Miletić, and S. Tschauner, 'Estimation of missing parameters for DICOM to 8-bit X-ray image export', in 2022 *International Conference on Electrical, Computer, Communications and Mechatronics Engineering* (ICECCME), Maldives, 2022.

[7] D. Barać, T. Manojlović, M. Napravnik, F. Hržić, M. Mamula Saračević, D. Miletić, and I. Štajduhar, 'Content-based medical image retrieval for medical radiology images', in 2024 *Artificial Intelligence in Medicine* (AIME). Lecture Notes in Computer Science, Cham: Springer Nature Switzerland, 2024, pp. 45–59.

[8] R. Baždarić, F. Hržić, M. Napravnik, and I. Štajduhar, 'Forming of validation dataset for deep learning based model of medical image grouping', in 2024 *Medical Imaging and Computer-Aided Diagnosis* (MICAD), Singapore: Springer Nature Singapore, 2023, pp. 411–429.

[9] M. Krajčí, M. Napravnik, and I. Štajduhar, 'Processing medical diagnostic reports using machine learning', in 2024 *47th MIPRO ICT and Electronics Convention* (MIPRO), Opatija, Croatia, 2024.