# Medical Image Retrieval for Augmenting Diagnostic Radiology

**Inauguraldissertation**

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät

der Universität Basel

von

Ashery Mbilinyi

Basel, 2023

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät
auf Antrag von

Prof. Dr. Heiko Schuldt, Erstbetreuer
Prof. Dr. Volker Roth, Zweitbetreuer
Prof. Dr. Henning Müller, Externer Experte

Basel, den 23.05.2023

Prof. Dr. Marcel Mayor, Dekan

*To my family*

*"There is no end to education. It is not that you read a book, pass an examination, and finish with education. The whole of life, from the moment you are born to the moment you die, is a process of learning."*

Jiddu Krishnamurti

# Abstract

Even though the use of medical imaging to diagnose patients is ubiquitous in clinical settings, their interpretations are still challenging for radiologists. Many factors make this interpretation task difficult, one of which is that medical images sometimes present subtle clues yet are crucial for diagnosis. Even worse, on the other hand, similar clues could indicate multiple diseases, making it challenging to figure out the definitive diagnoses. To help radiologists quickly and accurately interpret medical images, there is a need for a tool that can augment their diagnostic procedures and increase efficiency in their daily workflow. A general-purpose medical image retrieval system can be such a tool as it allows them to search and retrieve similar cases that are already diagnosed to make comparative analyses that would complement their diagnostic decisions. In this thesis, we contribute to developing such a system by proposing approaches to be integrated as modules of a single system, enabling it to handle various information needs of radiologists and thus augment their diagnostic processes during the interpretation of medical images.

We have mainly studied the following retrieval approaches to handle radiologists' different information needs; i) Retrieval Based on Contents, ii) Retrieval Based on Contents, Patients' Demographics, and Disease Predictions, and iii) Retrieval Based on Contents and Radiologists' Text Descriptions. For the first study, we aimed to find an effective feature representation method to distinguish medical images considering their semantics and modalities. To do that, we have experimented different representation techniques based on handcrafted methods (mainly texture features) and deep learning (deep features). Based on the experimental results, we propose an effective feature representation approach and deep learning architectures for learning and extracting medical image contents. For the second study, we present a multi-faceted method that complements image contents with patients' demographics and deep learning-based disease predictions, making it able to identify similar cases accurately considering the clinical context the radiologists seek.

For the last study, we propose a guided search method that integrates an image with a radiologist's text description to guide the retrieval process. This method guarantees that the retrieved images are suitable for the comparative analysis to confirm or rule out initial diagnoses (the differential diagnosis procedure). Furthermore, our method is based on a deep metric learning technique and is better than traditional content-based approaches that rely on only image features and, thus, sometimes retrieve insignificant random images.

# Acknowledgements

First and foremost, I would like to express my highest appreciation to my supervisor, Prof. Dr. Heiko Schuldt. I will forever appreciate your support and guidance throughout my Ph.D. journey. Thank you very much for believing in me and accepting me into your team. Thanks for the great discussions, feedback, and opportunities you have given me that have shaped me as a person and, most importantly, as a researcher. This dissertation would not be possible without you. Danke Sehr!

I would also like to sincerely thank Prof. Dr. Volker Roth for accepting me as my second supervisor. Moreover, I would like to thank Prof. Henning Müller for reviewing this dissertation as an external expert. Thank you very much for your time and commitment.

During my time as a member of the Databases and Information Systems Group (DBIS), I had the privilege of working with incredibly talented people, including Ivan Giangreco, Luca Rossetto, Lukas Probst, Mahnaz Parian-Scherb, Philipp Seidenschwarz, Marco Vogt, Ralph Gasser, Alexander Steimer, Dina Sayed, Shaban Shabani, Silvan Heller, Sein Coray, Loris Sauter, Florian Spiess, David Lengweiler and Rahel Arnold. Thank you very much for the great and insightful discussions. I have learned a lot from our conversations, and forever I will be grateful. I would also like to give special thanks to Yvonne Wegmuller for handling the logistics during my Ph.D. time.

I am lucky to have friends who supported me in one way or another throughout this journey. I wish to especially thank Tobias Schindler, Maria Karanatsios, Anneth-Mwasi Tumbo, Sammy Chebon, Carmen Thierstein, Daniella Bart, Harrison Macharia, Nadine Kugler, Debdeep Ghosal, Mathias Hebben, Xiao Meng, Arsenii Garamow, Katja Bennenberg, Flavia Mandini, Diep Nguyen, Olena Trush, and Anurag Arnab.

Lastly, I would like to thank my father, mother, sisters, and brothers. Thank you very much for your constant unconditional love and support. Ahsanteni sana!

This work was partly supported by the Swiss State Secretariat for Education, Research and Innovation in the context of a Swiss Government Excellence Scholarship for Foreign Scholars, which is also thankfully acknowledged.

# Contents

# List of Figures

# List of Tables

# List of Acronyms

| | |
|---|---|
| $MS_{loss}$ | Multi-Similarity Loss |
| AP | Average Precision |
| BERT | Bidirectional Encoder Representations from Transformers |
| CNN | Convolutional Neural Network |
| COVID-19 | Corona Virus Disease 2019 |
| CT | Computer Tomography |
| CXR | Chest X-Ray |
| DCG | Discounted Cumulative Gain |
| DenseNets | Densely Connected Convolution Networks |
| DML | Deep Metric Learning |
| EHR | Electronics Health Records |
| FC | Fully-Connected Layer |
| FDG | Fluorodeoxyglucose |
| GMU | Gated Multimodal Units |
| ICD | International Classification of Diseases |
| IR | Information Retrieval |
| LBP | Local Binary Pattern |
| MLP | Multilayer Perceptron |
| MRI | Magnetic Resonance Imaging |
| MRR | Mean Reciprocal Rank |
| NIH | National Institutes of Health |

| | |
|---|---|
| PET | Positron Emission Tomography |
| ReLU | Rectified Linear Unit |
| ResNets | Residual Neural Networks |
| RR | Reciprocal Rank |
| SIFT | Scale-Invariant Feature Transform |
| tf-idf | Term Frequency Inverse Document Frequency |
| US | Ultrasound |
| WHO | World Health Organization |

Part I

# Introduction

# 1

# Introduction

In December 2019, the WHO was informed about the disease outbreak with an unknown cause in Wuhan, China [JXY+20]. Among early patients were a 49-year-old woman (in the following, called patient one) and a 61-year-old man (in the following, called patient two). Both were admitted to a hospital on December 27, 2019, with fever, cough, and chest discomfort [ZZW+20]. Patient one was diagnosed with Pneumonia, a lung inflammatory condition commonly affecting tiny air sacs of the lungs (alveoli) [Pne22]. This diagnosis was through the findings observed by a CT imaging examination. Pneumonia was also detected in patient two, and its manifestations were through a CXR imaging examination. Both patients underwent pneumonia treatment, including an oxygen ventilator to help them breathe. Patient one was recovered and discharged from the hospital on January 16, 2020, while patient two died on January 9, 2020 [ZZW+20].

The cause of the disease in both patients was later identified as a new virus, also known as a novel coronavirus (SARS-CoV-2) [ZZW+20; LGW+20]. On $11^{th}$ of February 2020, the disease was officially named COVID-19 and declared a global pandemic on $11^{th}$ of March 2020 by WHO [WHO22b]. Since then, COVID-19 has become among more than eighty thousand diseases/conditions present in the WHO's International Classification of Diseases (ICD), a global standard of diagnostic health information [WHO22a; ICD22]. COVID-19 has brought about more than 632 million cases and 6.59 million deaths, as of $16^{th}$ of November 2022 (see Figure 1.1), making it one of the deadliest in history [TFE22a; WHO22c].

Even though COVID-19 was a new disease then, this story is an excellent example that showcases two critical things:

- First, it shows how vital medical imaging is as a diagnostic tool to understand what is wrong underneath the human body.

- Second, it shows how important it is for clinicians to quickly and accurately in-

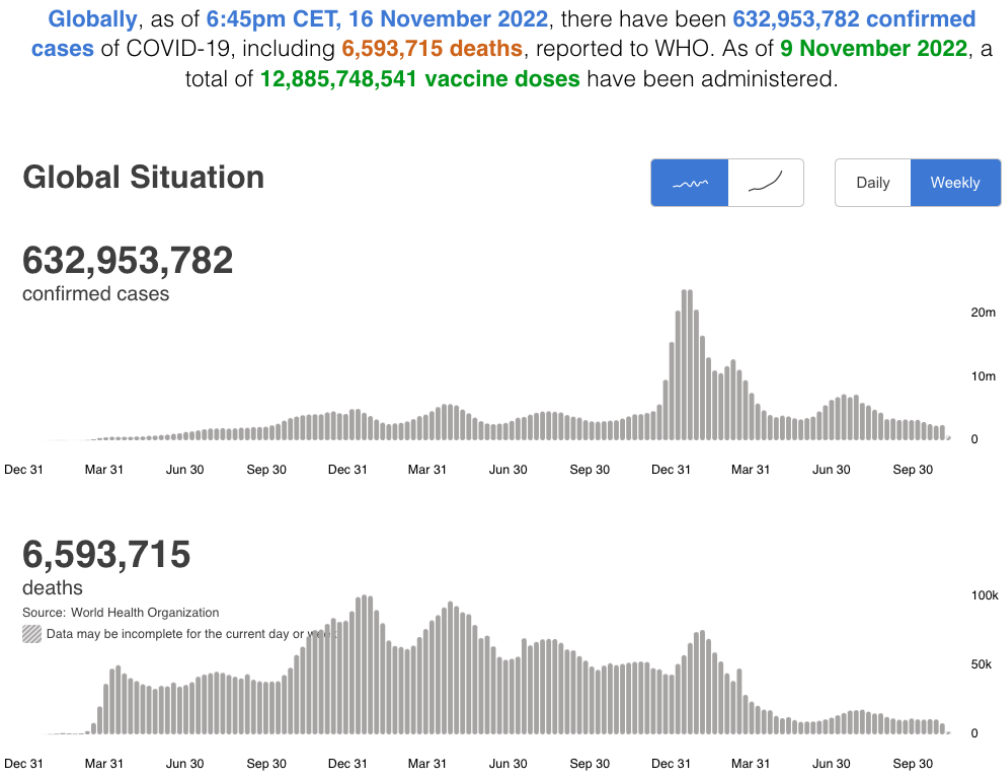


**Figure 1.1 WHO COVID-19 Dashboard [WHO22c].**

terpret medical images, as the outcome can make a difference in a patient's life.

In everyday clinical environments, patients typically go through a diagnostic procedure that is not that different from those mentioned above COVID-19 patients went through. Usually, a patient would first visit a Medical Doctor for consultation after feeling unwell. Depending on the patient's case, a medical doctor would order various tests, including a medical imaging examination, to diagnose a patient. Medical imaging examination is the preferred method because it allows clinicians to investigate the happenings inside human bodies without surgery or other invasive procedures. It is one of the best ways to diagnose patients with no harmful side effects.

Most medical imaging examinations fall under the Diagnostic Radiology umbrella. By definition, diagnostic radiology is a branch of medicine relying on non-invasive imaging scans for patients' diagnosis. Diagnostic radiology can diagnose many problems, including heart conditions, gastrointestinal conditions, broken bones, blood clots, etc. In contrast, Interventional Radiology, another type of radiology, uses imaging technology to help guide medical procedures, e.g., treating cancers, blockage in arteries or veins (Angiography), liver and kidney problems, e.t.c. [Cli22]. A clinician specializing in diagnosing and treating diseases using medical imaging as their fundamental tool [WGA+19] is known as a Radiologist. Typically a radiologist's workflow follows the steps shown

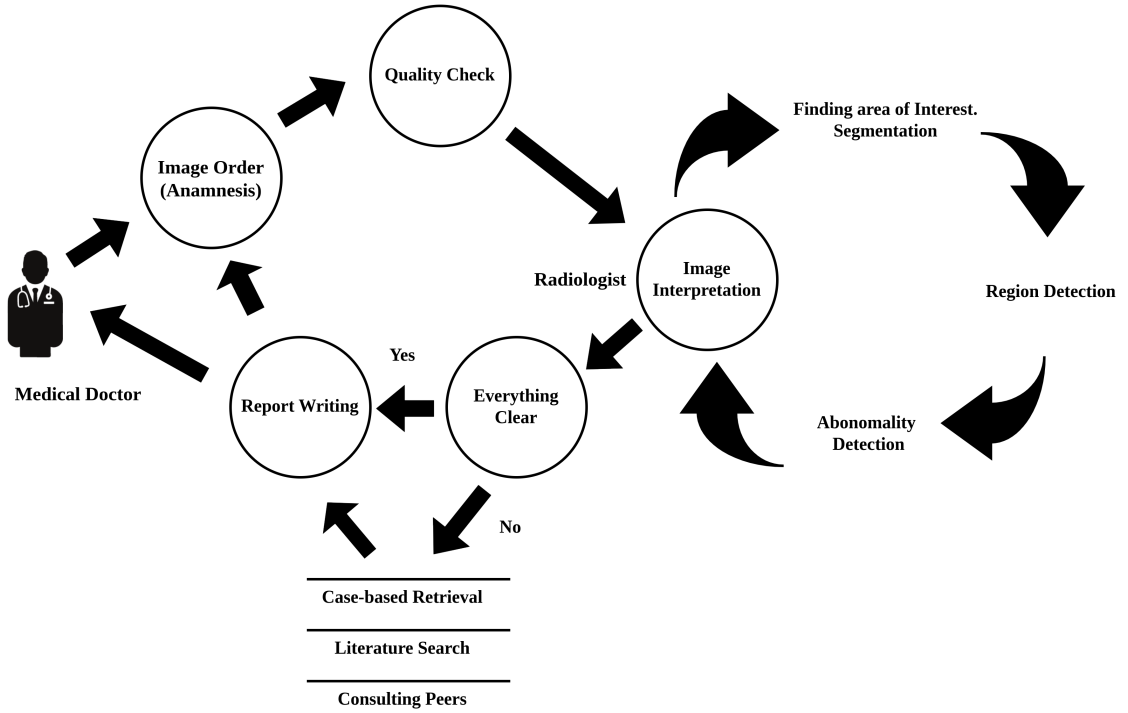


**Figure 1.2 The radiologist's workflow example in a clinical environment (Adapted from [MÜ17]).**

in Figure 1.2, in which following an order for the imaging examination by a medical doctor, a radiologist would first acquire an image using imaging technology like X-rays, CT e.t.c and then perform a detailed quality check of the image. Afterward, they would interpret the image and compile a report detailing their findings [MÜ17], and a medical doctor would use these reports, among other things, to determine the treatment a patient needs.

## 1.1 Motivation

Unfortunately, the radiologist's workflow when interpreting medical images is not straightforward. Instead, it is an iterative and challenging process often prone to errors. Many factors contribute to this situation; among them is the subjective nature of diagnostic radiology [WGA+19] itself.

An excellent example of this subjectivity is the so-called *availability bias*, a tendency that immediate past experiences easily influence diagnostic assessments. Itri et.al [ITM+18] observed this phenomenon in which a radiologist diagnosed a circled region of interest in a CT image (seen in Figure 1.3) as perigastric hematoma instead of lymphoma simply because he encountered a case shown in Figure 1.4 that had similar visual patterns but had perigastric hematoma earlier on the same day.

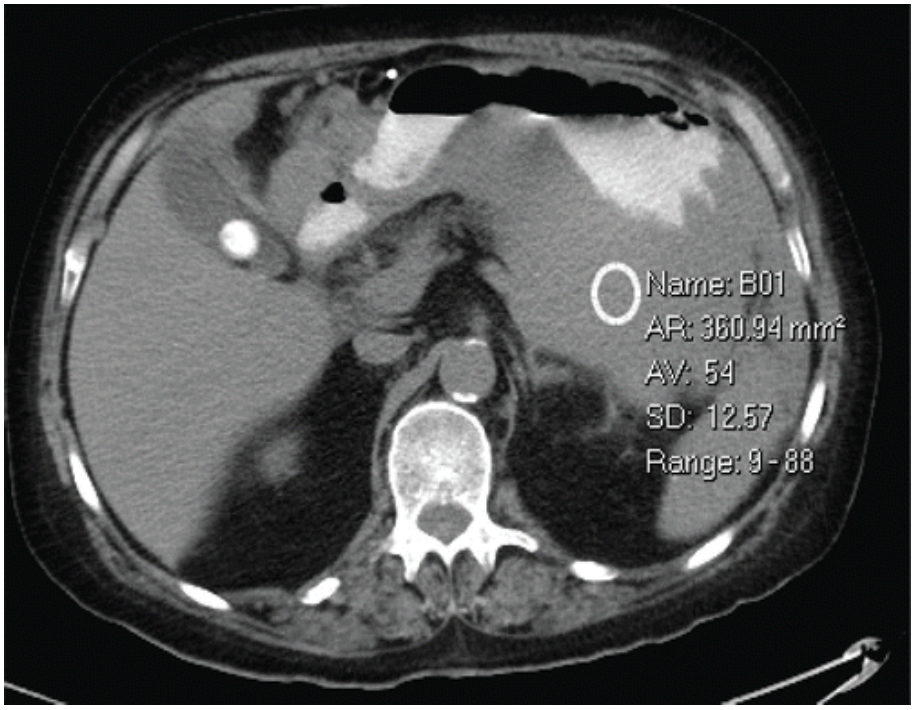


**Figure 1.3 Nonenhanced CT image shows a perigastric area of attenuation (circled region of interest) [ITM⁺18].**

Subjectivity is also contributed by perception, which is the most critical skill in diagnostic radiology. In general, the perceptual expertise of radiologists is defined by their refined visual search patterns [KRD⁺16], which highly correlates with their years of experience. Expert radiologists not only perceive abnormalities that non-experts do not, but they also better understand what to attend to and ignore [GP19]. This means not only that an intern radiologist is likelier to make errors than a specialist, but even a specialist is prone to make diagnostic errors when they encounter a novel case.

Studies have shown that the estimated rate of diagnostic errors in radiology ranges between 3% to 5%, and there are forty million diagnostic errors involving imaging annually worldwide [ITM⁺18]. In the United States alone, nearly 5% of adults who seek outpatient care are likely to experience diagnostic errors. On the other hand, diagnostic errors contribute to up to 17% of adverse hospital events [SMT14] and nearly 10% of all deaths annually. These errors are costly since they result in wasteful medical spending. An annual estimated cost typically ranges between $17 billion to $29 billion [DCK⁺00]. From the legal side, around 75% of malpractice lawsuits against radiologists are also due to their errors in diagnosis [Ber96; BCG18].

To eliminate diagnostic errors, Itri et al. [ITM⁺18] proposed *a swiss cheese model of safety* that offers many layers of defense to eliminate different factors contributing to the likelihood of making such errors. One component in these layers includes providing clinical decision support that augments radiologists' workflow and helps them make

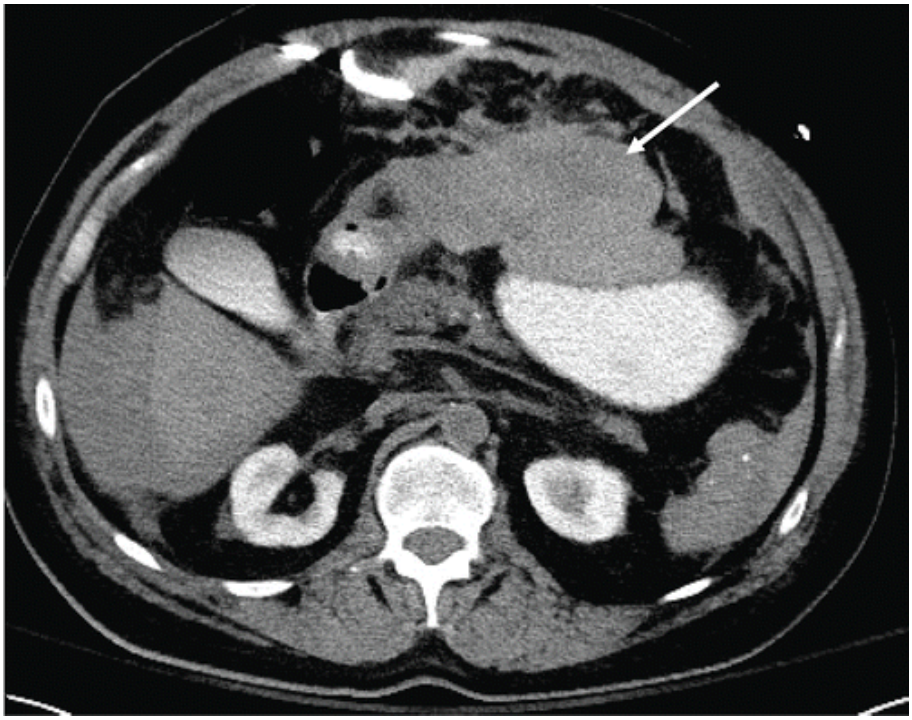

**Figure 1.4 Axial contrast-enhanced CT image showing a perigastric hematoma (arrow) [ITM+18].**

accurate clinical decisions. A medical image retrieval system can provide such support. An excellent example of its use-case is when a novice radiologist wants to interpret a CXR image with local white opaque patches, indicating either consolidation or pneumonia [RZK+19]. To confirm or rule out these initial diagnoses, he/she wants to retrieve confirmed similar cases from the hospital archive to do a comparative analysis that can showcase the similarity or dissimilarity of those cases with the image at hand. Hence, he/she can get a shred of evidence needed to arrive at the definitive diagnosis.

However, accurately identifying and retrieving similar cases takes more than just analyzing image contents alone. A medical image retrieval system needs to consider clinical contexts, which means it needs to analyze information of different types, like the demographic of the patients, symptoms, laboratory data, e.t.c. to accurately determine the similarities between different cases, just like how a radiologist would need to consider information from different sources to diagnose and compare patient cases accurately [HPS+20]. On the other hand, we need to consider that radiologists are trained professionals. When they read medical images to interpret them, they form initial hypotheses on the likely diagnoses before eventually arriving at the definitive diagnoses. This means that when they want to retrieve similar images, sometimes they already know what kind of images they need to search for so that the retrieved image can add value to the comparison analysis needed to form definitive diagnoses. From a medical image retrieval system perspective, a system needs to take as input both images radiol-

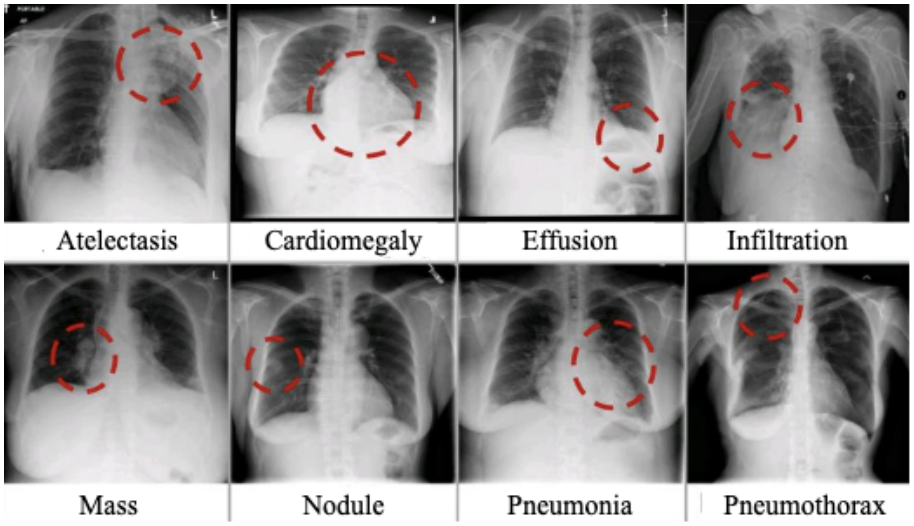


**Figure 1.5 Thoracic diseases in CXR images as seen by a radiologist [WPL+17].**

ogists want to compare and their initial hypotheses to help perform a targeted retrieval per radiologists' information needs.

## 1.2 Information Need

When radiologists interpret medical images, they do so with the background knowledge they acquired from their medical training. This means they can infer abnormalities that may indicate the presence of certain diseases/conditions, something a layperson cannot do when looking at similar images. This information deduced by the radiologist's eyes is semantic and differs from how a computer interprets medical images. Examples in Figures 1.5 and 1.6 illustrate this phenomenon. The former shows how radiologists spot abnormalities (red circles) that indicate thoracic diseases in CXR images. The latter shows how computers would see similar images one by one.

Computer sees images as a matrix of numbers (between 0 to 255), also called pixels, the smallest units in an image. Each pixel contains a different number of channels. When images are grayscale, they would have only one channel, but if they were colored images, they would contain three channels: red, green, and blue [Pok19]. This difference between how radiologists and computers interprets medical images is generally known as a *semantic gap* and has enormous implications for how radiologists search for medical images. To illustrate more on this, assume a radiologist named John is interpreting a patient's frontal and lateral CXR images with multiple densities (abnormal whiteness) in both lungs, as shown by black arrows in Figure 1.7. As John keeps gazing, he finds that the larger densities are ill-defined, and maybe there is an air-bronchogram in the right lower lobe. These abnormalities could hypothesize that either the patient has multifocal

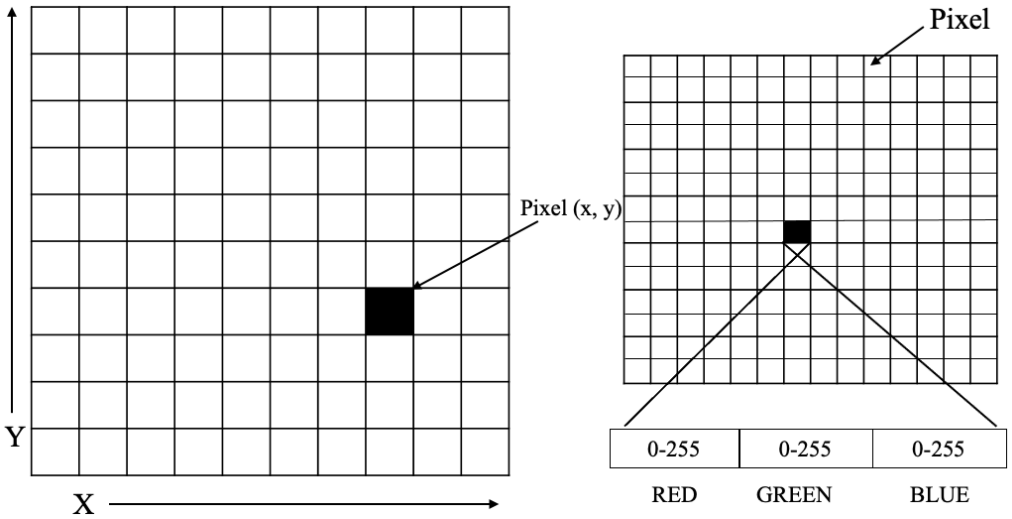


**Figure 1.6 Illustration of image pixels as seen by a computer [Pok19].**

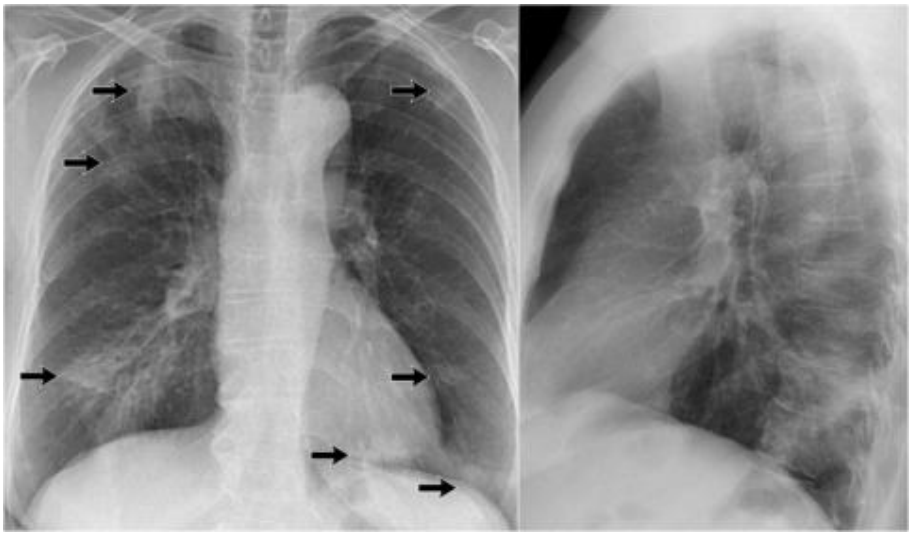

**Figure 1.7 Frontal (left) and lateral (right) CXR images with multiple densities (marked with black arrows) [Smi22].**

consolidations or multiple ill-defined masses [Smi22].

To perform the differential diagnosis, a process of ruling out or confirming initial hypothetical diagnoses, John wants to search for images from a hospital archive with similar visual patterns but have either of the initial hypotheses so he can compare them with the image he is trying to interpret. In this scenario, it would be easier for John to get the results that satisfy his information needs if a computer could infer the semantic concepts in medical images like disease labels instead of pixel values.

Luckily, the rise of deep learning, which is the study of models involving a composition of either learned concepts or learned functions [GBC16d] in a large amount, has the potential to reduce this semantic gap. Through deep learning, a computer can learn the abstract semantics of medical images by learning their representations through a nested

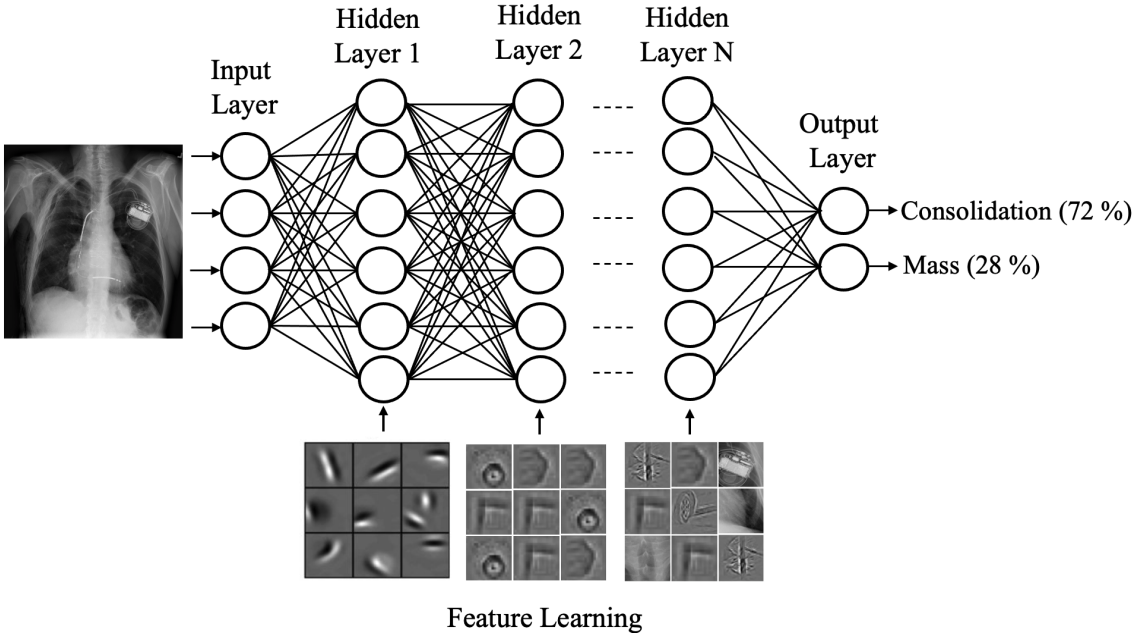


**Figure 1.8 A deep learning model illustration.**

composite of concepts, with each concept defined in connection to basic concepts and more abstract representations computed as a function of less abstract ones. This allows computers to map input raw data, like a collection of pixel values, to high-level semantic information, such as disease labels, as shown in Figure 1.8 [GBC16d]. Due to this capacity of deep learning, we have heavily relied on deep learning techniques to inform the retrieval processes in this thesis.

## 1.3 Contributions

In the big picture, we envision a general-purpose medical image retrieval system that can handle various radiologists' information needs. Such a system should be flexible enough to accommodate radiologists' different ways of expressing such needs and deliver the required results accurately. Developing such a system, however, is an ambitious task that needs many moving pieces brought together. It needs collaboration between stakeholders, including computer scientists and clinicians, especially in collecting and annotating the required medical images and the other patients' data for accurate diagnosis and identification of similarities between different cases.

This thesis contributes to developing such a system through three studies we conducted in chronological order in which the approach of retrieving medical images in a subsequent study is influenced by the limitations learned from the approach of the previous study. On the other hand, these studies were also conducted depending on the datasets available to experiment with the proposed methods. Our studies can be

summarised as follows:

- In the first study, we looked into the basic medical image retrieval approach, retrieval based on contents. Here a radiologist expresses information need by submitting a sample image (query by example), and the system computes the similarities of medical images based on their contents. To accurately retrieve the required images, we aimed to find an effective feature representation method to identify similarities between medical images considering their semantics and modalities.We have experimented with different representation techniques based on deep learning (deep features) and handcrafted techniques (mainly texture features). The experimental results show that deep features are superior to handcrafted features. It also shows which deep learning architectures can effectively learn and extract medical image visual contents. On the other hand, the study shows the limitation of relying on content alone, as sometimes the system retrieves insignificant random images, which does not add value in augmenting radiologists' diagnosis workflow.

- Owing to the lesson of the first study, our second study aimed at supplementing medical image contents with patient demographics meaning a radiologist can express information needs by submitting a query comprising a sample image and patient's demographic information. On the other hand, using deep learning, we added disease predictions to further helps in the identification of similarity between different cases. By leveraging images' content features, patient demographics, and disease predictions, our method can understand the clinical context and thus accurately improve the identification and retrieval of similar cases. Even though this method improves the retrieval results, it is limited in helping radiologists with specific comparison analyses like the one needed during the differential diagnosis procedure. To augment a differential diagnosis process, radiologists would want to retrieve specific images that are significant for the particular analysis.

- In the third study, we propose a method that allows radiologists to target specific images to retrieve. In this method, radiologists can submit queries consisting of sample images and text descriptions expressing their information needs. Based on a deep metric learning technique, our proposed method then learns how to combine the image contents and text descriptions to form a query and also the embedding function that puts the formed query closer to the targeted images and further from other images in the embedding space. By leveraging image contents

and text descriptions, our method guarantees that radiologists retrieve significant images for the comparative analysis needed for differential diagnosis procedures.

As mentioned above, these three studies were limited to the availability of the required public datasets to validate and prototype the proposed methods. Nevertheless, their methods can be extended to different medical images and patient data when available. In the following, we briefly outline the contributions of this thesis:

- We present a detailed empirical study of different feature representation techniques based on handcrafted methods (mainly texture features) and deep learning (deep features) to retrieve medical images.

- We propose an effective feature representation method and deep learning architectures for learning and extracting medical image features.

- We propose a method that accurately identifies and retrieves similar medical image cases considering their clinical context by leveraging image contents, patient demographics, and deep-learning-based disease predictions.

- We propose a guided search method that combines image contents with radiologists' text descriptions to target retrieving specific images suitable for the comparative analyses needed during the differential diagnosis procedures.

## 1.4 List of Publications

The following papers have resulted from some of the work presented in this thesis:

- Ashery Mbilinyi and Heiko Schuldt. Cheres: a deep learning-based multi-faceted system for similarity search of chest x-rays. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, pages 669–676, 2022.

- Ashery Mbilinyi and Heiko Schuldt. Retrieving chest x-rays for differential diagnosis: a deep metric learning approach. In *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–4. IEEE, 2021.

- Ashery Mbilinyi and Heiko Schuldt. Cross-modality medical image retrieval with deep features. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2632–2639. IEEE, 2020.

## 1.5 Thesis Structure

This thesis is organized into four parts. The current part motivates and introduces the problem while the subsequent parts are as follows:

- **Part II** introduces the fundamental concepts applied in this thesis, starting with the brief introduction of Diagnostic Radiology in Chapter 2. While Chapter 3, 4 and 5 covers Deep Learning Fundamentals, Medical Image Retrieval Systems, and Similarity in Medical Images, respectively.

- **Part III** presents three studies conducted in this thesis, namely i) Content-Based Retrieval (Chapter 6) ii) Retrieval Based on Contents and Disease Predictions (Chapter 7) and iii) Retrieval Based on Contents and Radiologists' Text Descriptions (Chapter 8). Each chapter highlights the problem studied and the methodology used, including the experiment results, and finally proposes an approach to address such a problem and how that approach contributes to developing a general-purpose medical image retrieval system.

- **Part IV** finalize the thesis with Conclusion in Chapter 9 while Chapter 10 outline future directions of this research.

Part II

# Foundations

# 2

# Diagnostic Radiology

Understanding how different medical imaging technologies work sheds light on the challenges for radiologists and computer systems in interpreting images produced by these technologies. This chapter briefly introduces diagnostic radiology and the standard medical imaging technologies also known as Medical Imaging Modalities. We further explain the principles behind these modalities in producing images and what information they are examining when diagnosing the human body.

## 2.1 Introduction

Diagnostic radiology, also called diagnostic imaging or medical imaging, is the branch of medicine that relies on non-invasive imaging scans for diagnosing patients [Cli22]. Due to the diversity of human body tissues and structures of various organs and their diseases and conditions, different diagnostic radiology technologies exist so far. On the other hand, scientists continue inventing different technologies to find the proper mechanism to examine the internal structure of the human body to get the correct diagnoses. Currently, some conventional medical imaging technologies include, Radiography also known as X-rays, Ultrasound, and Nuclear Medicine, also known as Gamma Cameras. Others include Magnetic Resonance Imaging, Computed Tomography, Diffusion Tensor Imaging, Optical Coherence Tomography, and, Positron Emission Tomography [Goa20]. Both technologies generally facilitate the working human body's internal aspects to simulate the 2-D or 3-D digital images for medical examinations [Kit22].

## 2.2 Medical Imaging Modalities

Medical imaging modality refers to the technique and process used to visualize a particular part of the body, organs, or tissues for diagnostic purposes [Kit22]. Each modality

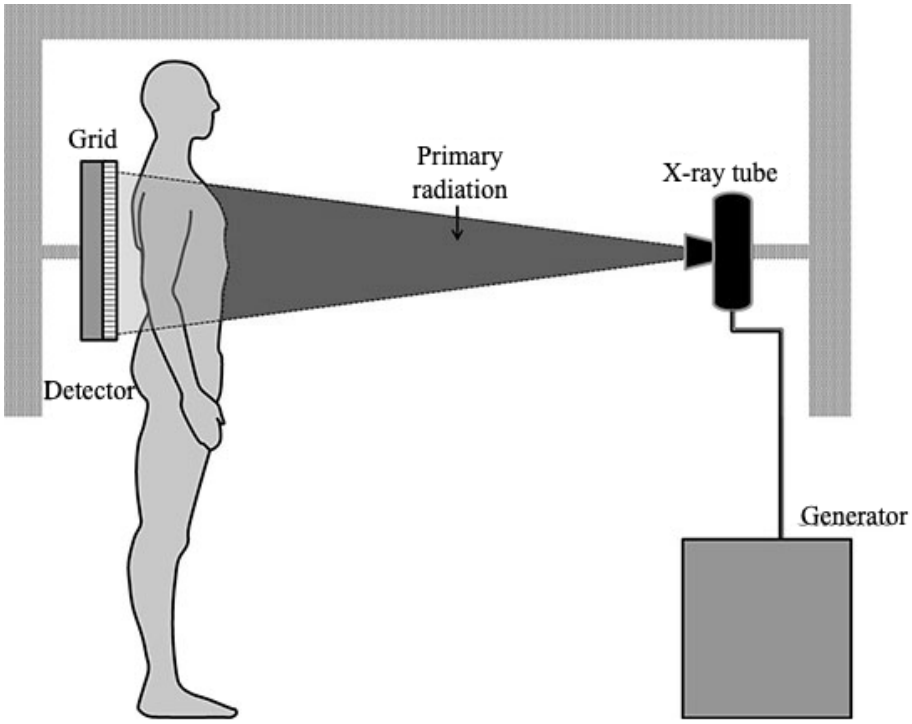


**Figure 2.1 An illustration of X-ray imaging procedure [Key23].**

usually exploits a physical phenomenon such as radioactivity, magnetic resonance, electromagnetic radiation, the propagation of sound waves, etc., to generate digital images or visualizations of the human body's internal tissues or a part of the human body in a manner that is non-invasive [ZGD+20]. Said differently, each imaging modality maps a specific physical parameter [Goa20]. The information produced by this mapping is what a radiologist uses to infer abnormalities when diagnosing a patient. Therefore, the technology and machinery used in each imaging modality vary depending on which physical parameters a radiologist wants to measure. Some of these technologies measure a particular physical parameter using radiation while others do not [Cli22].

The following sections briefly explain the mechanics behind conventional modalities (Radiography, Computed Tomography, Positron Emission Tomography, Ultrasound and Magnetic Resonance Imaging, and ) with respect to the physical parameters they are measuring.

### 2.2.1 Radiography

Diagnostic radiography relies on X-ray radiations, also known as Röntgen rays, to examine the body. These radiations were accidentally discovered by Wilhelm Conrad Röntgen when he found a glow of crystals near a high-voltage cathode-ray tube while working at Würzburg University, Germany. He then concluded that the cathode-ray tube generated energy that could penetrate the nearby paper, causing the crystals to glow [Kit22].

**Figure 2.2 The X-ray images for different body parts [CDC22].**

Since no one understood these fascinating rays, he named them X-rays, a notion from the mathematical designation for something unknown. In other words, X-rays are unknown radiations. Typically, an X-ray image maps how different tissues absorb different amounts of radiation. Human bodies comprise substances of varying densities, and hence, by highlighting these differences using the absorption or attenuation of X-ray photons, the X-ray film can reveal the internal structure of the body [Cat22]. To produce X-ray images, the X-ray machine's tube is usually directed toward the patient's body area of interest, e.g., Chest, Knee, etc. X-ray radiations from the tube are then passed through the body, in which a radiation detector on both sides of the body processes the radiations to visualize the interiors. The process is entirely done without incisions, as illustrated in Figure 2.1. The examples of X-ray images produced for different body parts can be seen in Figure 2.2. Typically, bones usually appear white since the calcium within absorbs X-ray radiation the most. On the other hand, fat and other soft tissues would look grey since they absorb less, while air absorbs the least, so the lungs usually appear black. Any abnormalities in how these tissues absorb the radiation give a diagnostic clue to radiologists. Diagnostic radiography, commonly referred to as

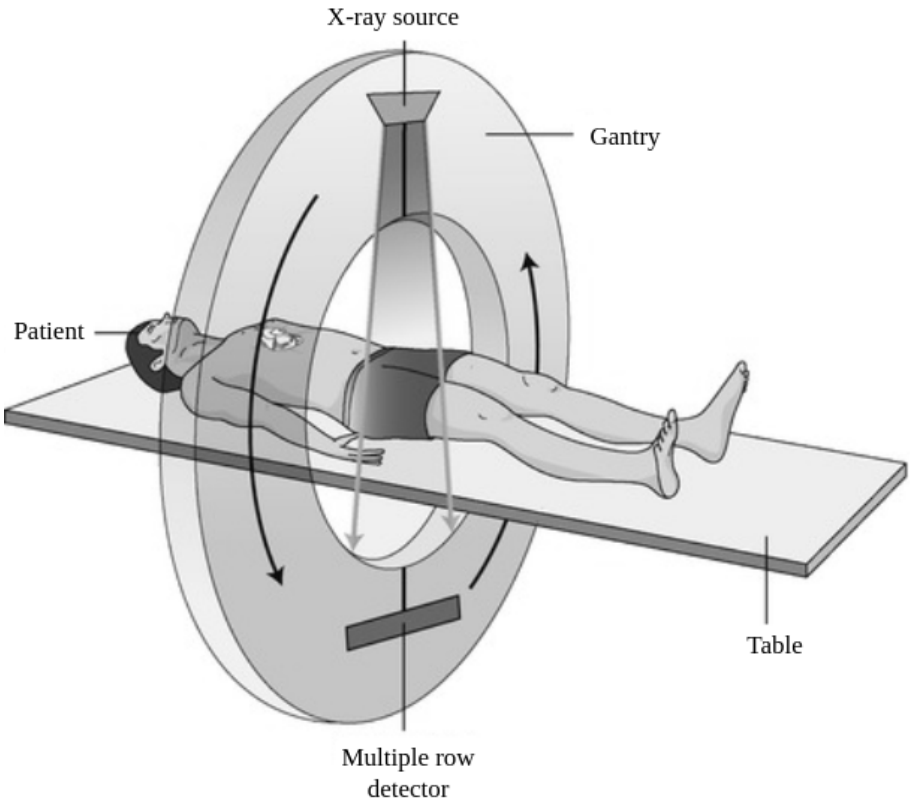


**Figure 2.3 An illustration of the CT imaging procedure [Tho23].**

X-ray, comes second after laboratory tests as the most widely used medical test, [Cat22]. It remains one of the most critical examinations globally for screening, diagnosis, and managing various diseases that are life-threatening [IRK+19].

### 2.2.2 Computed Tomography

The main drawback of radiography is that the information's depth is entirely lost because the film produces images representing the X-ray beams' total attenuation when they pass through the patient [Cat22]. To address this drawback, Godfrey Hounsfield initially thought by considering X-ray readings of all angles around the object enclosed in the box, one could figure out the depth information. To illustrate this concept further, Godfrey assembled a computer to acquire position points by focusing X-rays at different angles to construct an image of the hidden object. Subsequently, based on this concept, he developed the prototype for the application in medicine that saw the first clinical CT scan performed on a patient with a frontal lobe tumor at St. George's Hospital in London, the UK, in 1972 [Kit22]. Just like Hounsfield's initial idea, the process of producing a CT image involves taking many X-ray images of the region of interest by letting a patient lie on the bed that moves to and fro towards a rotating ring as Figure 2.3 illustrates. The amount of X-ray radiation absorbed relates to the thickness of the slice and is propor-

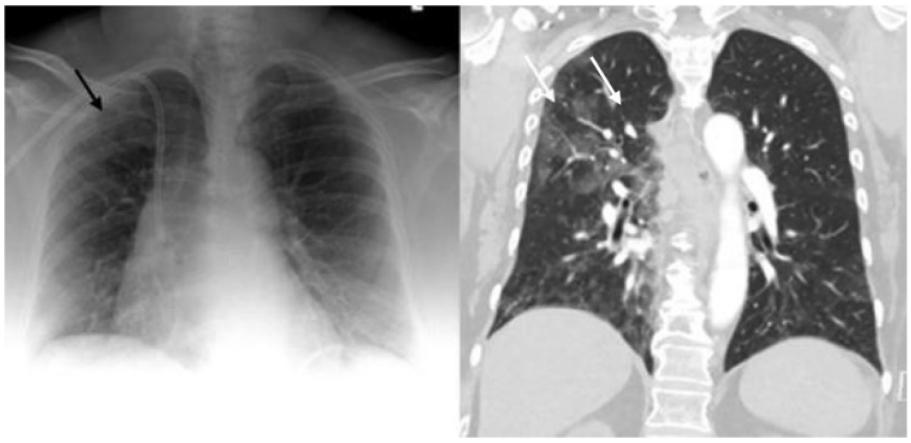

**Figure 2.4 A contrast between chest X-ray (left) and CT (right) images [JCB+20].**

tional to the tissue's density. Therefore, by changing the angle of the X-ray source tube, a computer can correlate the tissue densities into a cross-sectional image. A CT scan is mainly used by radiologists in need of highly detailed images in order to figure out the course of a problem, especially on soft tissues like small nodules or tumors, which is difficult to obtain with a standard X-ray machine. Sometimes, depending on the body organs to be examined, a patient must take a contrast agent that makes the body absorb X-ray radiations enough to make some structures more evident in the generated images. Figure 2.4 illustrates differences between X-ray and CT images by contrasting a chest X-ray and CT images of the same patient. Figure 2.5 shows a high resolution, low dose chest CT with the tracheal stent.

### 2.2.3 Positron Emission Tomography

A PET scanner relies on a small dose of radiation to check the activity of cells in various parts of the body. It gives more in-depth details about cancer or abnormal parts shown using CT scans, X-rays, or magnetic resonance images [Mac22]. In other words, it tells more than what X-rays, CT scans, and magnetic resonance images can show about cancer or other abnormalities. On the surface, the PET imaging procedure looks similar to CT imaging procedure, but the physics behind the machine is different. A patient is first injected with a radioactive tracer called Fluorodeoxyglucose (FDG) through the arm. The PET scanner then tracks the radiation emitted by the tracer over a specific time frame to create images indicative of radiation levels accumulating inside the body. Said differently, a physical parameter the PET scanner measures is the number of co-incident gamma rays emitted due to the decay of radiotracer as shown in Figure 2.6 [Goa20]. FDG is close to glucose, and the cancer cells will absorb this radiotracer much faster compared to normal cells [Kit22]. Accordingly, if no FDG levels are indicated, that would mean a normal body functioning of the particular body region. Figure 2.7 shows a contrast be-

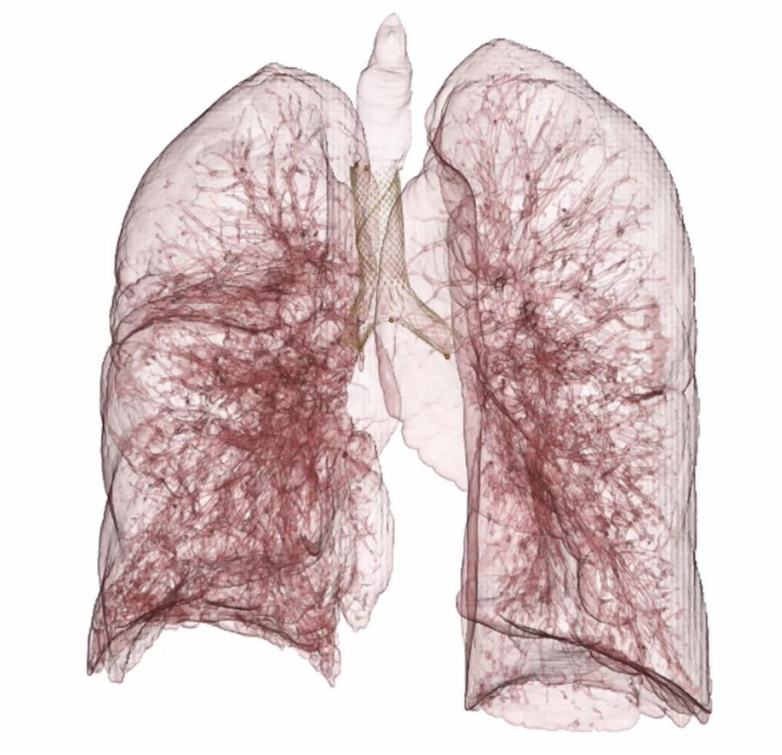

**Figure 2.5 High resolution, low dose chest CT with tracheal stent [Goa20].**

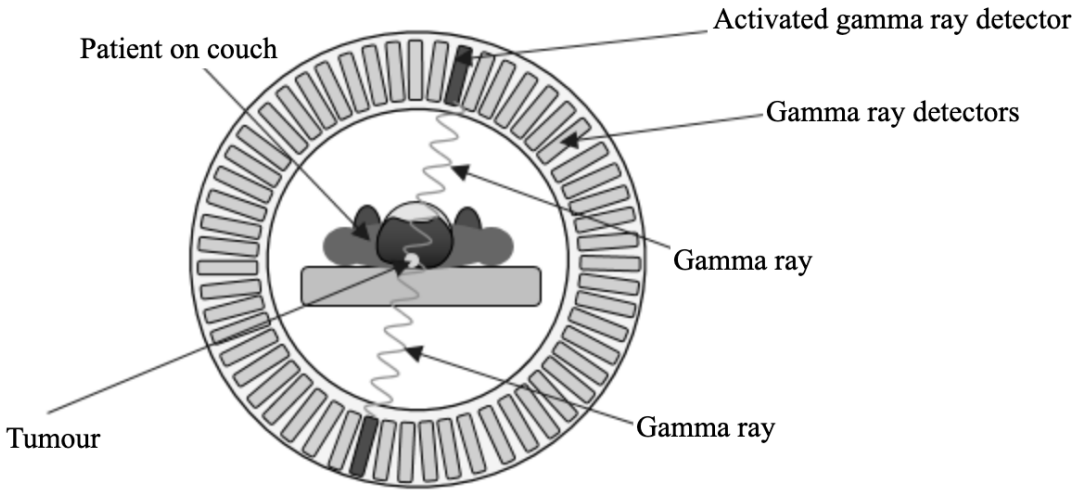


**Figure 2.6 An illustration of a PET imaging procedure [Goa20].**

tween a CT and PET image of a 68-year-old male with lung cancer [KKY+16].

### 2.2.4 Ultrasound

The medical US took inspiration from underwater sonar research on using echo to locate and identify objects. Therefore, unlike the previously mentioned medical imaging modalities that use radiation, Ultrasound, on the other hand, relies on sound waves to

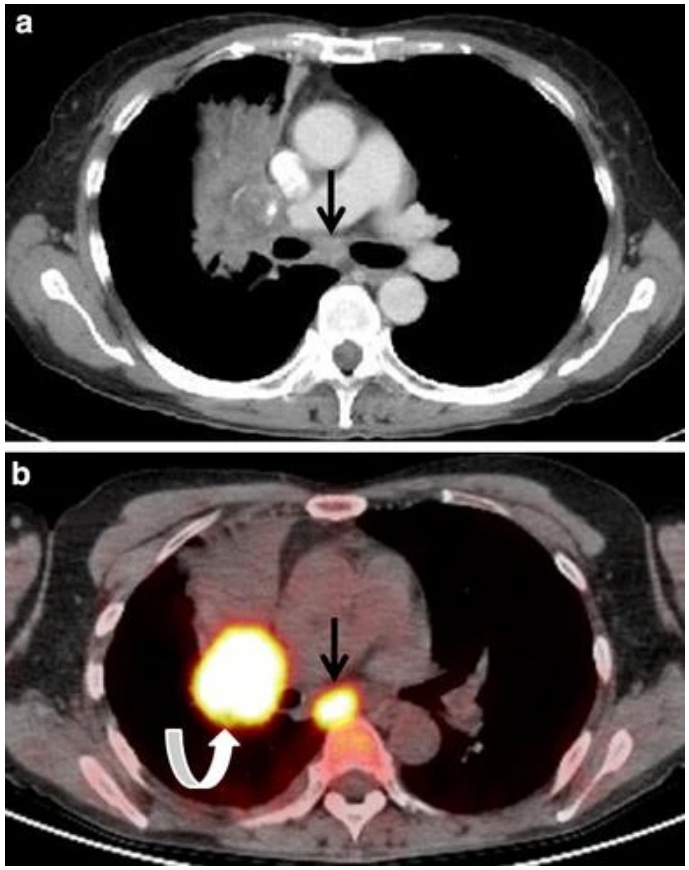


**Figure 2.7 Comparison between CT (a) and PET (b) images. Arrows within the images illustrate the difference in lung cancer manifestations between these two modalities [KKY+16].**

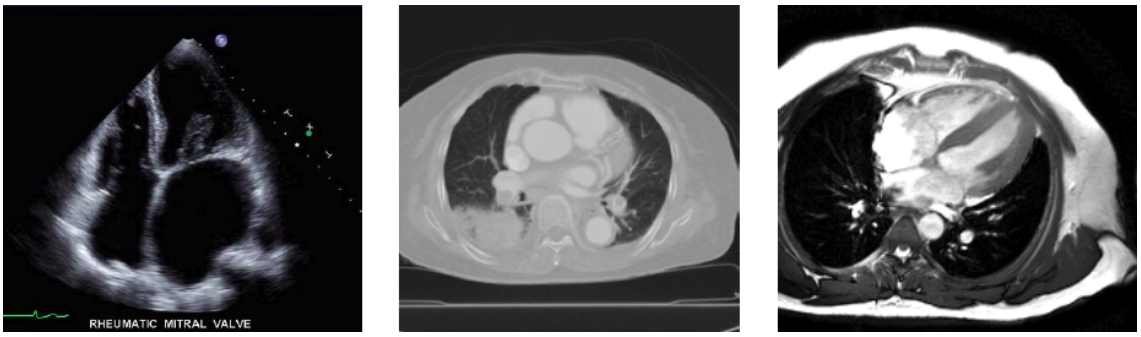


**Figure 2.8 Comparison of US (left), CT (middle), and MRI (right) [Cat22].**

construct a visualization of the human's interior organs.

Using a transducer capable of emitting high-frequency sound waves (greater than 1 MHz, which is above the threshold of human hearing) [Goa20], Ultrasound scanners pass these waves into the human body. The sound waves reflect off different organs and their surrounding tissues to make echoes, which bounce back to the transducer. When these echoes arrive at the transducer, they construct electrical signals that are then transmitted to the Signal Processor. Using the speed of sound and the time of each echo's arrival, the

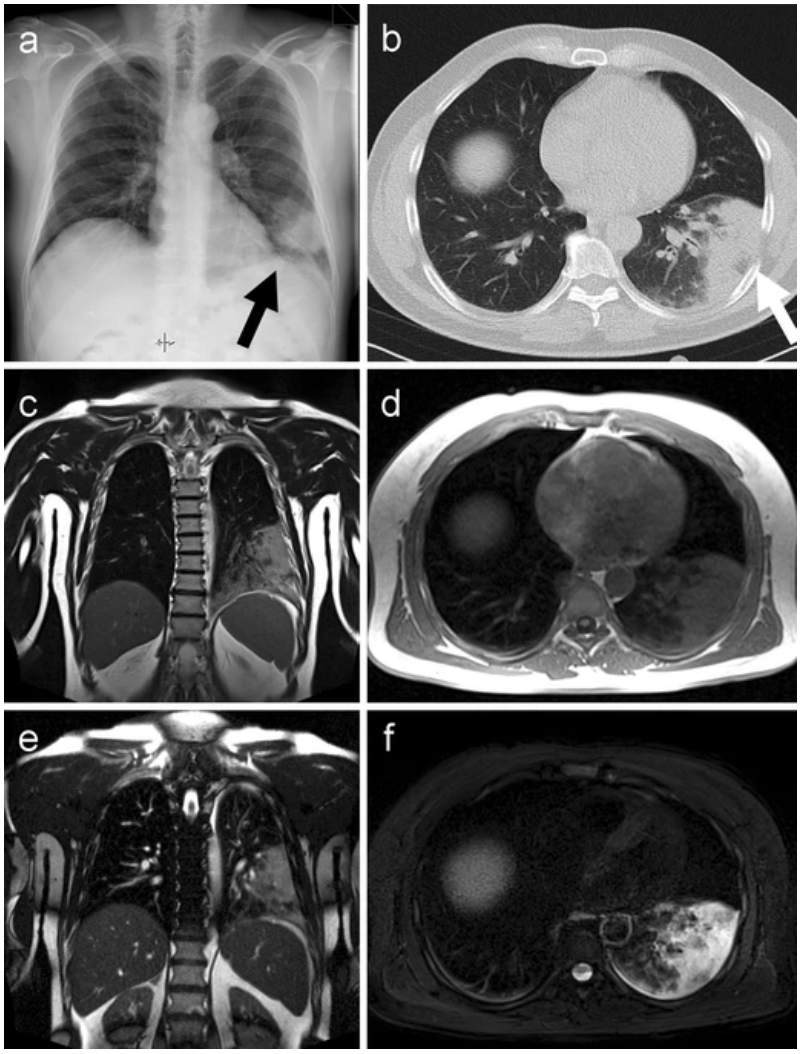


**Figure 2.9 Comparison between CXR (a), CT (b), and MRI images (c,d,e, and f). Arrow within the images illustrates the difference in the manifestations of pneumonia [BBH+12].**

signal processor computes the distance between the transducer and the tissue boundary. This distance information is then used to construct two-dimensional images of organs and tissues [BIB22b]. In other words, the ultrasound image maps the backscattering of sound waves as a physical parameter [Goa20]. Ultrasound can investigate various problems in the human body, including gastrointestinal lung diseases, abdominal or chest pain, cardiology, etc. Ultrasound is also mostly used to get real-time images of unborn children [Kit22]. Since US images tissue boundaries instead of, e.g., density information [Cat22], the images produced are slightly different from other modalities. Figure 2.8 shows a comparison between US, CT and MRI.

### 2.2.5 Magnetic Resonance Imaging

Magnetic resonance imaging modality uses magnetism to generate a detailed image of body areas. The human body comprises 63% hydrogen atoms from its water and fat components [Kit22]. MRI is based on a complex technology that excites and detects the change in the orientation of the rotational axis of protons that are in these water that form living tissues [BIB22a]. It is suitable for examining soft tissues or non-bony body parts, like the brain, muscles, nerves, spinal cord, tendons, and ligaments.

The magnetic resonance imaging procedure involves subjecting a patient to an intense magnetic field and a series of radio waves. The MRI scanners employ powerful magnets that produce an intense magnetic field that aligns the body's protons with it [BIB22a]. When a radiofrequency current is then pulsed through the patient, the protons are excited and spin out of the equilibrium, forced against the pull of the magnetic field. When the radiofrequency field is turned off, the MRI sensors can detect the energy that comes out as the protons, the changes in the quantity of energy released depending on the environment and the chemical nature of the molecules, and the duration it takes for the protons to realign with the magnetic field. Radiologists can infer the difference between different types of tissues as the function of these magnetic properties. The MRI imaging procedure is similar to the CT (see Figure 2.3); however, MRI usually takes longer, is much nosier, and requires the patient to stay still during the whole procedure. Figure 2.9 shows a contrast between a CXR, CT and MRI images of a 66-year-old male with Pneumonia [BBH+12].

## 2.3 Summary

In this chapter, we have briefly introduced the field of diagnostic radiology and different imaging modalities used to diagnose patients. We have also explained the physics behind each modality and the physical parameter measured as visualized by the images produced. With a trained eyes view, radiologists usually interpret these medical images with respect to the physical parameters measured, and therefore they can infer high-level semantic information. As explained in Section 1.2, since the computer interprets images using low-level features only, a semantic gap between what a radiologist and a computer see is created. Deep learning can help to reduce this gap. In the next chapter, we introduce the fundamentals of deep learning and explain how deep learning techniques help reduce this semantic gap.

# 3

# Deep Learning Fundamentals

With the help of deep learning, a computer system can map raw data input, like the collection of pixel values, to high-level semantic information, like disease predictions. This reduces the semantic gap in interpreting medical images between radiologists and computer systems, making deep learning a critical tool for informing medical image retrieval systems. In this chapter, we review the fundamentals of deep learning, starting with a brief introduction to deep learning in general and then diving into different types of neural networks used in this thesis.

## 3.1 Introduction

Deep learning is a particular kind of machine learning that achieves greater power and flexibility by representing information as a nested hierarchy of concepts, with each concept built up from more straightforward concepts and more abstract representations computed in terms of less abstract ones [GBC16d]. The illustration in Figure 1.8 shows an excellent example of this. Here, a deep learning model learns to represent the concepts within a CXR image by integrating simpler concepts like contours and corners to an abstract concept like the probability of diseases diagnosed. As explained in Section 1.2, there is a semantic gap between what a radiologist sees in a medical image (see Figure 1.5) as compared to how a computer would see the same image (see Figure 1.6). By learning a more abstract representation of medical images, a deep learning model can help to reduce this semantic gap and therefore helps to identify the similarity between medical images during the retrieval process. In the abstract, when developing a deep learning model, one must first understand the nature of the task and how to design the appropriate model. For the former, most deep learning problems fall into either supervised or unsupervised tasks. In supervised learning, usually, there is a ground truth, and the model has to learn the underlying approximation function that maps the raw input

data to the ground truth information.

In contrast, an unsupervised learning task does not need the ground truth. Here, a model's task is to recognize underlying patterns in the raw input data. For the latter, designing a model is not a straightforward process; instead, one must consider many hyper-parameters. Among these hyper-parameters, the neural network architecture is one of the critical design parameters. In the subsequent section, we introduce the neural network architectures we used in this thesis, starting with a multi-layer perceptron, the straightforward and typical neural network architecture.

## 3.2 A Multilayer Perceptron

The goal of Multilayer Perceptron (MLP) is to approximate some function $f^*$. For example, for a classifier, $y = f^*(x)$ maps an input $x$ to a category $y$. An MLP defines a mapping $y = f(x;\theta)$ and learns the value of the parameter $\theta$ that makes the best function approximation [GBC16c]. In other words, a multilayer perceptron is a mathematical function mapping input and output values [GBC16d]. These input values are passed through a unit known as a neuron, designed based on the inspiration of how the core building block of the brain works.

The neurons are assembled in a chain-like structure to form a network of neurons (hence the name *neural networks*) to create a function composing many simpler functions. For example, we might have two functions $f^{(1)}$ and $f^{(2)}$ connected in a chain, to form $f(x) = f^{(2)}(f^{(1)}(x))$. These chain structures are the most widely used neural network structures. In this case, $f^{(1)}$ is called the first layer of the network, $f^{(2)}$ is called the second layer, and so on. The chain's overall length produces the model's depth; hence the name *deep learning* [GBC16c]. Said differently, deep learning refers to the use of an artificial neural network composed of a large number of layers [Fon20]. Moreover the deeper the neural network architecture, the more it can potentially lead to progressively more abstract representations at higher layers [BCV13]. Figure 3.1 illustrates an example of an MLP with three layers: an input layer, hidden layer and output layer.

Like any learning algorithm, an MLP learns $\theta$ by adjusting the value of weights and biases for each input to produce the desired output through a learning process that requires four components, *a dataset, a model itself, a loss function, and an optimization strategy* [GBC16e]. The dataset's quality largely determines how well the model learns the optimal weights. On the other hand, designing the model includes choosing the correct hyperparameters like the neural network architecture, including how many layers the network should compose, how these layers should be structured, and how many neurons should be in each layer [GBC16c]. In the following, we briefly describe the

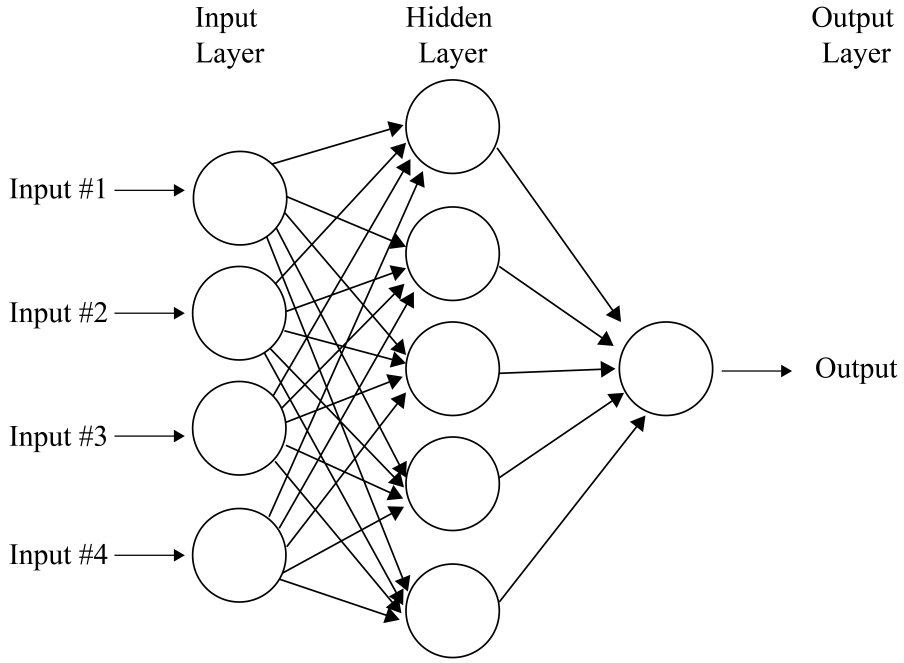


**Figure 3.1 An example of the MLP with three layers [MNZ+15].**

remaining two components, a loss function, and an optimizer.

**Loss function:** A loss function is the most critical aspect of designing a deep learning model [GBC16c]. Given the input, the neural network flows information forward to produce the predicted output. The loss function helps to tell the gap between the predicted and target values. In turn, the model learns the optimal weights to produce the desired predictions by minimizing the loss function with the help of the backpropagation algorithm [RDG+95]. The backpropagation algorithm does that by allowing information from the loss to flow backward through the network to calculate the gradients which informs the model on how to adjust the network weights in each training iteration. In other words, any loss function must be differentiable.

**Optimizer:** In the abstract, an optimizer task helps the model finds the parameter $\theta$ that significantly reduces a loss function $L(\theta)$. Since neural networks primarily use nonlinear activation functions, which makes the optimization nonconvex, this is a problem of finding the global minimum. It is difficult, however, to find the global minimum in the nonconvex function. Luckily for large deep neural networks, most local minima have a minimal loss function value; therefore, a genuinely global minimum is unimportant [GBC16f]. This means most optimizers in neural networks find a parameter $\theta$ after finding good-enough local minima. Many optimization algorithms exist; however, the fundamental and often used algorithm is Gradient Descent [Rud16].

In theory (the universal approximation theorem [HSW89]), any large MLP can learn to approximate the desired function. However, learning may fail for two reasons. First,

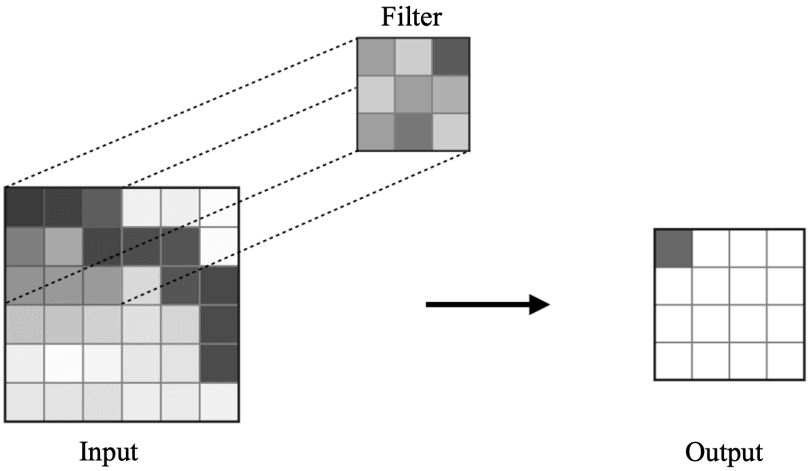


**Figure 3.2 The Convolution and Pooling Operations [AA18].**

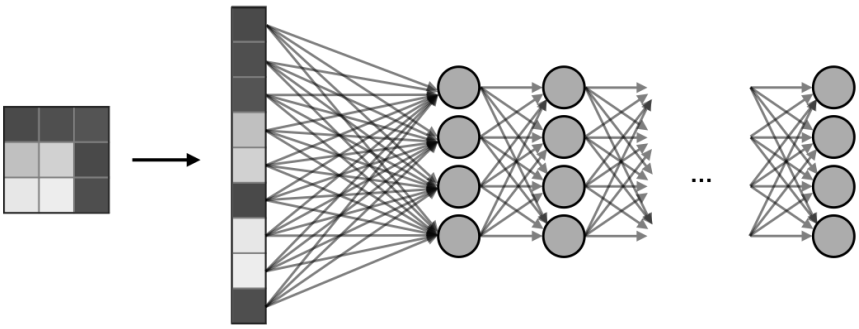


**Figure 3.3 Fully Connected Layers [AA18].**

the optimization algorithm used for training may not find the value of the parameters suitable for the target function. Second, the training algorithm could select the wrong function because of overfitting [GBC16c]. One way to overcome these shortcomings of MLPs is to design a task-specific neural network architecture that can approximate the desired function. In the next section, we briefly introduce Convolutional Networks, a particular neural network designed to learn appropriate functions for computer vision tasks.

## 3.3 Convolutional Neural Networks

Convolutional Neural Network (CNN), also known as Convolutional Networks, ConvNets, or CNNs, are specialized kinds of neural networks suitable for processing data that are in the multiple arrays form, for example, an image composed of three 2D arrays containing pixel intensity values [GBC16b; LBH15]. These networks learn the high-level semantics from raw input signals by exploiting the property that many natural signals are compositional hierarchies in nature, in which higher-level features are acquired by

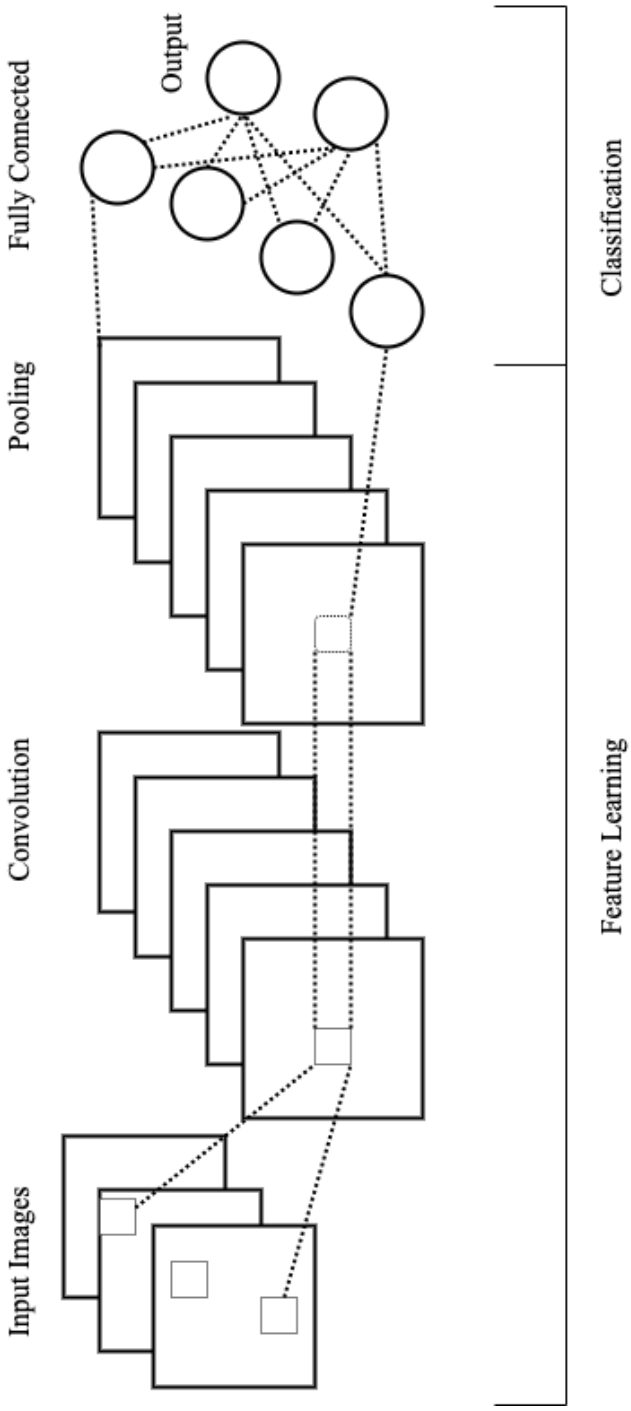


**Figure 3.4 A general framework of convolutional neural networks [LZM+18].**

integrating lower-level features [LBH15]. In images, local combinations of edges form motifs, motifs combine into parts, and parts create objects. This property allows CNNs to learn high-level abstract representations of images suitable for most computer vision tasks, including image retrieval.

CNNs combines four key ideas to learn image features effectively: shared weights, local connections, pooling, and the use of many layers to take advantage of the properties of raw input signals [LBH15]. These layers are of three types: convolutional, pooling, and the fully-connected. The convolution layers use filters to perform convolution operations when passing through the input $I$ with respect to its dimensions. Its hyper-parameters involve the size of the filter $F$ and stride $S$. The resulting output $O$ is a*feature map* also known as *activation map* [AA18]. The pooling layer performs a down-sampling operation, usually applied after a convolution layer, to deal with spatial invariance (see Figure 3.2). Particulary, average, and max pooling are particular kinds of pooling where the average and maximum values are considered, respectively. Fully-Connected Layer (FC), on the other hand, deals with a flattened input whereby each input is connected to all neurons (see Figure 3.3). If present, FC layers are typically at the end of CNNs and are used to optimize specific tasks like class score predictions.

In other words, the general framework of convolutional neural networks involves the feature learning part, which comprises convolutional and pooling layers, and the task-specific part, which involves fully connected layers as Figure 3.4 shows [LZM+18] where the task-specific part is for the classification task.

Even though CNNs have already been very successful in many computer vision tasks and applied in many commercial applications [GBC16b], the optimal way to design their architectures to effectively learn image features is still an open question for researchers to date. In the following, we briefly overview some of the most common and successful convolutional networks' architectures.

**AlexNet:** AlexNet [KSH17] won the 2012 ImageNet object recognition challenge which, as a result of that success, the vast interest in both commercial and research of deep learning arose [GBC16b]. As shown in Figure 3.5, this model's architecture contains eight learned layers which are five convolutional and three FC layers.

**VGG-16:** Simonyan and Zisserman [SZ14] developed this model for the 2014 ImageNet Challenge, in which it took the first position in the localization and a second position in the classification tasks. The idea behind this architecture was to investigate the effect of the depth of the CNN on its accuracy in large-scale image recognition. After thoroughly evaluating networks of increasing depth using tiny $(3 \times 3)$ filters, they concluded that a critical improvement on the prior-art configurations could be achieved by making the depth of the network between 16-19 layers, hence the name VGG16. Fig-

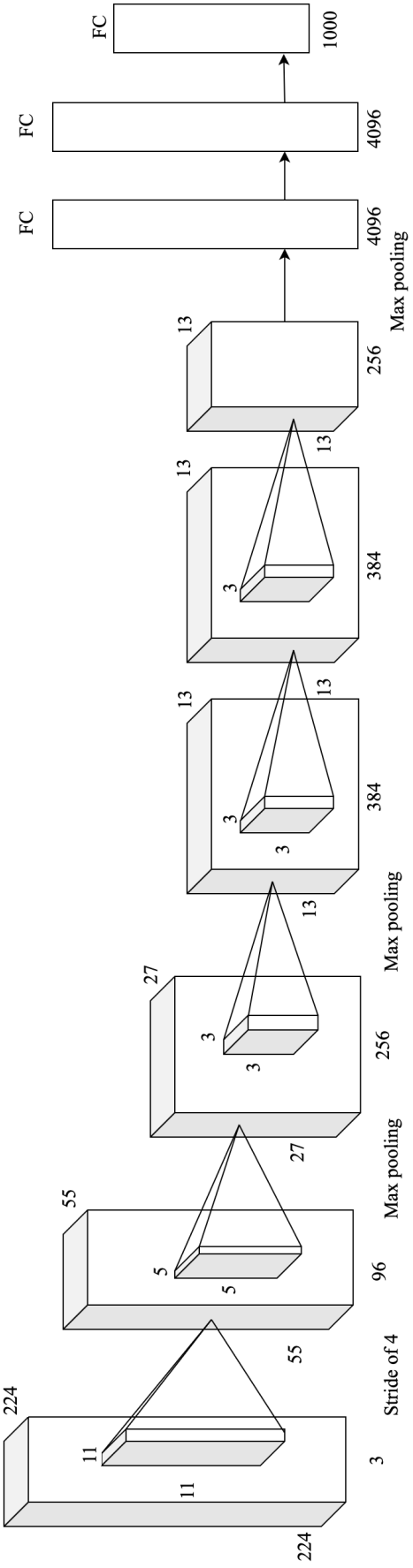


**Figure 3.5 AlexNet architecture [KSH17].**

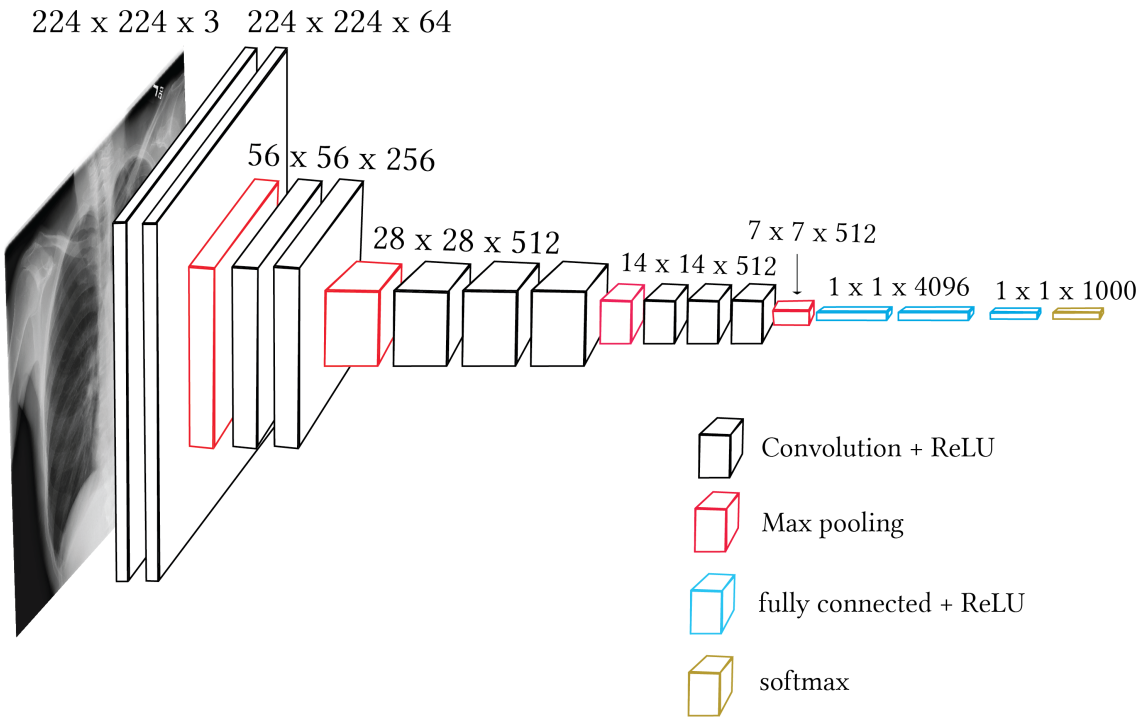


**Figure 3.6 VGG16 architecture [SZ14].**

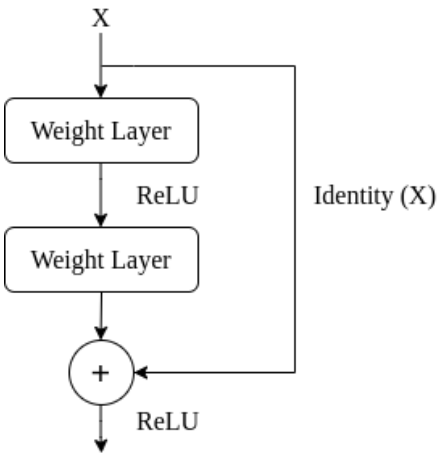


**Figure 3.7 A residual learning block [HZR+16].**

ure 3.6 illustrates the VGG16 architecture.

**ResNets:** Residual Neural Networks (ResNets) provides deep representation, which is of core significance for various visual recognition tasks, and for that matter, it won first place in the 2015 ImageNet localization and classification challenges. This architecture presented a residual learning framework to soothe the training of deeper networks [HZR+16], hence the name Residual Networks. To deal with the degradation problem, He et al. introduced residual blocks in which intermediate block layers learn a residual function with relation to the block input. This is like a refinement step in which

**Figure 3.8 Example of a residual network with 34 layers in contrast with 34 layers plain network and VGG-19 network architecture [HZR+16].**

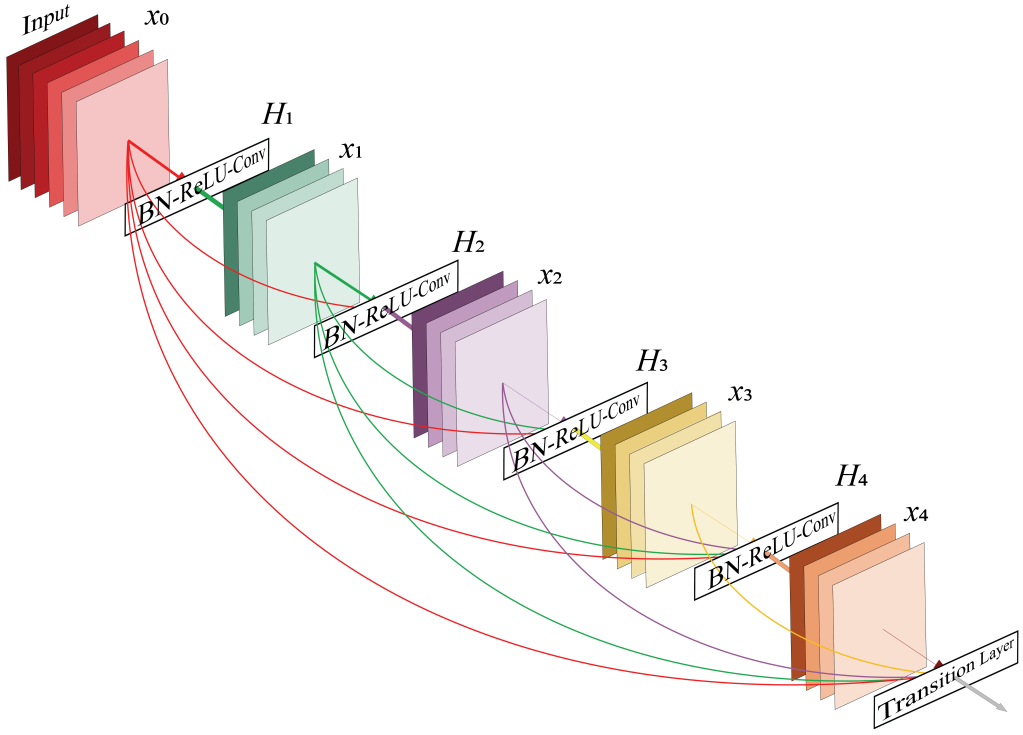


**Figure 3.9 A 5-layer dense block [HLM+17].**

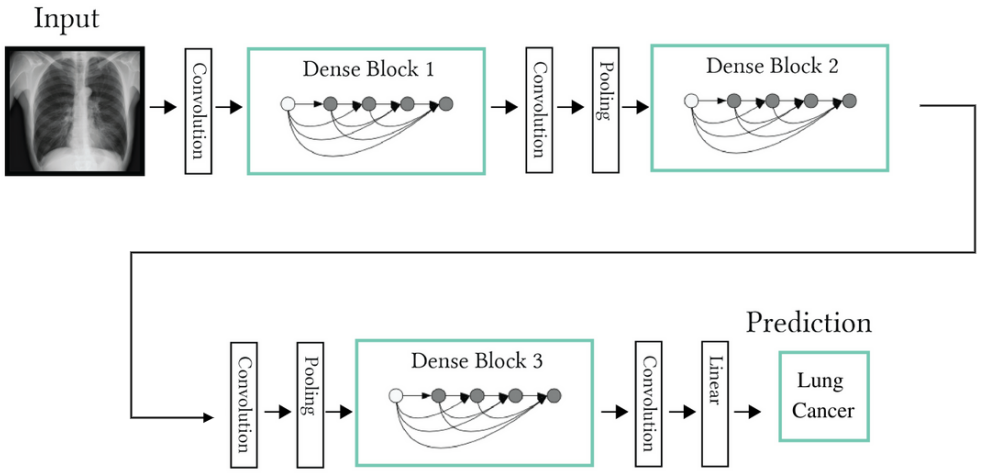


**Figure 3.10 A DenseNet with three dense blocks [HLM+17].**

the network learns to regulate the input feature map for higher-quality features, which is not similar to a "plain" network architecture in which each layer learns new and different feature maps. If the refinement is unnecessary, ResNets intermediate layers can learn to gradually change their weights toward zero so that the residual block represents an identity function. Figure 3.7 illustrates the building block for residual learning in s while Figure 3.8 shows a contrast of a  with 34 layers compared to a plain and VGG-19 networks architectures.

**DenseNets:** A DenseNet architecture is a variant of ResNets, but instead of having

a residual block, Huang et al. [HLM+17] introduced a dense block as shown in an example of a five-layer dense block in Figure 3.9. Within a dense block, the feature maps of all proceeding layers are used as inputs for each layer. Similarly, its feature maps are also used as inputs into all following layers. This concatenation of feature maps learned by various layers improves the variation in the information of subsequent layers and, therefore, enhances efficiency. In turn, Densely Connected Convolution Networks (DenseNets) provide robust feature propagation and promote the reuse of features.

As we previously explained, finding an optimal network architecture is still an open research question. Nevertheless, finding its solution can shed light on a better way to learn effective high-level feature representation, hence an optimal feature extractor for image features, including medical images. However, even though deep convolutional networks can learn abstract concepts for medical images, they still produce representations of higher dimensions. This introduces a higher computational cost for similarity search needed for image retrieval tasks. The following section presents a neural network architecture specifically for dimensionality reduction, the autoencoder.

## 3.4 Autoencoders

An autoencoder is the kind of neural network trained to attempt to duplicate its input to its output [GBC16a]. This network has a hidden layer $h$, internally that describes a *code* that represents the input. Typically, the network consists of two parts: an encoder function ($h = f(x)$) and a decoder that does the reconstruction ($r = g(h)$) (see Figure 3.11).

Given an input image $x$, the encoder takes $x \in R^d = X$ and maps it to $h \in R^p = F$, i.e., $h = f(x)$ where $h$ is the internal representation or *code* of $x$ in a feature space $F$ with lower dimension compared to input space $X$ (i.e., $d > p$). Conversely, the decoder $g$ reconstructs the original image $x$ by mapping its representation to the reconstructed image $r$. Learning is done by minimizing the reconstruction loss $L(x, r) = ||x - r||$. By representing the input into a lower dimensional feature vector, the autoencoder prioritizes essential aspects of the input to be duplicated; therefore, it usually learns critical properties of the raw input data.

In other words, an autoencoder can learn high-level semantic information from raw input data by selecting only abstract concepts from the raw input data.

This means it is not especially useful if an autoencoder simply learns to set $g(f(x)) = x$ everywhere. Therefore, autoencoders are designed to not successfully learn to duplicate input data perfectly [GBC16a] and hence only learn essential abstract concepts. Autoencoder is the quintessential example of a representation learning algorithm [GBC16d]

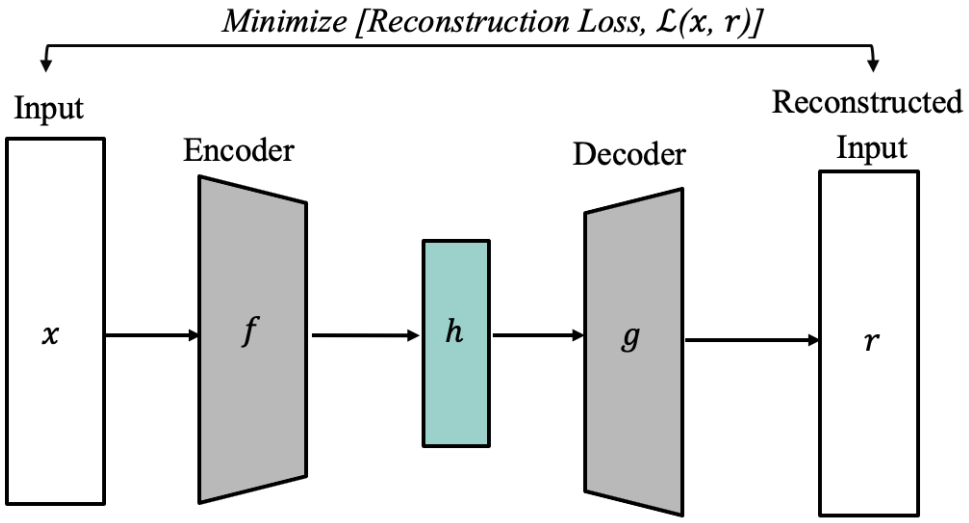


**Figure 3.11 A general structure of an autoencoder, mapping an input $x$ to an output (called reconstruction) $r$ through an internal representation (code $h$). The autoencoder comprises two components: the encoder $f$ (mapping $x$ to $h$) and the decoder $g$ (mapping $h$ to $r$) [GBC16a].**

and has been successfully applied for information retrieval and dimensionality reduction tasks [GBC16a].

## 3.5 Summary

This chapter briefly introduced deep learning and explained different neural network architectures, including MLPs, CNNs, and autoencoders. We detailed how these neural network architectures learn the feature representations of images to encode high-level semantic information. Since deep neural networks can reduce the semantic gap in the interpretation of medical images between radiologists and computer systems, they are critical tools for informing the retrieval process in medical image retrieval systems. In the next chapter, we will discuss conventional medical image retrieval systems approaches especially their feature representation techniques.

# 4

# Medical Image Retrieval Systems

Medical image retrieval systems as a research field fall under Information Retrieval, in which three retrieval approaches, text-based, content-based, and multi-modal, are conventional. In these three approaches, feature representation is the most critical part of the system as it determines how the similarities between different cases of medical images are identified. In other words, feature representation is the prerequisite to achieving excellent performance in any medical image retrieval system [LZM+18]. In this chapter, we briefly introduce Information Retrieval and then deep dive into medical image retrieval systems, specifically into feature representation techniques. Finally, we conclude with an overview of the evaluations of the retrieval systems.

## 4.1 Introduction

Information Retrieval (IR) is the process of finding material of an unstructured nature through extensive collections (usually stored on computers) for the purpose of satisfying a particular information need [MRS10]. Typically, an IR system searches in collections of unstructured or semi-structured data (e.g., documents, images, web pages, video, etc.) [SC12] as opposed to structured information as dealt with by a database management systems (DBMS) [Cre00].

The general conceptualization of the information retrieval process can be formalized as follows [WTK+21]:

1) The user has an information need and forms a query;

2) The user submits the query to the IR system;

3) The IR system ranks the results based on a similarity function and returns the top-ranked results;

4) The user checks the results and decides to continue or stop.

This process depends on the assumption that there is a feature space where the representations for both the user's queries and the data available in the collection can be compared. Therefore the general retrieval model can be expressed as a tuple as shown in 4.1.

**Definition 4.1 Generic retrieval model** (based on [Gia18, p. 34]).

A retrieval model is a tuple

$$[\mathcal{D}, \mathcal{Q}, \mathcal{F}, \delta(\cdot, \cdot)]$$

where

- $\mathcal{D}$ denotes a set of representations of data in a collection;
- $\mathcal{Q}$ denotes a set of representations of user needs through the query;
- $\mathcal{F}$ denotes a framework to model the data representations and the queries with a feature transformation function $f(\cdot) \in \mathcal{F}$;
- $\delta(\cdot, \cdot)$ is a comparison function, given as a similarity function or a distance function.

A medical image retrieval system is no different from a typical IR system in which a user has to send a query ($\mathcal{Q}$) to get results of top-ranked images evaluated by their relevance (as computed by $\delta(\cdot, \cdot)$) by to the query. Traditionally, the standard way to form a query has been to issue a text expressing information needs. The retrieval of such a query would depend on if the texts in the query match the metadata of images available in the archive. However, this text-based approach is tricky because it assumes accurate annotations of medical images exist, which is not always the case in clinical settings.

Another way of expressing the information need popular in the literature has been a *"query by example"*. Here a user, a radiologist in this context, sends a sample image, and the retrieval system searches for similar images by comparing the contents of the sample image to images available in the collection. Said differently, this is a content-based approach. While this approach is widely preferable for medical images as it alleviates the errors that come with unreliable annotations, it still has drawbacks. As explained in Section 1.2, a semantic gap exists between how a computer sees medical images and how a radiologist's trained eyes interpret the same images. This means the notion of similarity between images as evaluated by a computer system might vary from how a

radiologist would judge. To address this gap, the feature representations of images need to accommodate high-level semantic concepts similar to how a radiologist interprets images. Said differently, computers must learn to see images as a radiologist does.

## 4.2 Text-Based Retrieval

Text-based medical image retrieval systems can be traced back to the 1970s, and they are prevalent in the search on the internet web browsers [HLC12]. In these systems, users usually express their information needs through a text query, and the retrieval results are based on matching the query text and the metadata associated with particular images available in the archive. In other words, the retrieval efficiency depends on i) how the system figures the user's intent while processing the text query. ii) if the information from the texts matches the annotations available in the archive.

The matching is usually done by employing two standard techniques, namely *Boolean retrieval* and *vector-space retrieval*. In Boolean retrieval, the system checks if the query's exact or intended words are also in the metadata with the help of boolean operators *AND, OR* and *NOT* that guide the evaluation of how relevant the respective metadata is to the query. In other words, referring to Definition 4.1, boolean operators act as comparison functions to check the similarity between the text query ($\mathcal{Q}$) and annotations ($\mathcal{D}$). This retrieval type is also known as exact-match retrieval and does not rank the retrieved results.

On the other hand, in vector-space retrieval, also known as partial-match retrieval, the query terms are further processed to form a vector representation. This vector representation is then compared to the metadata representations to determine their similarities through algebraic operations such as dot product or calculating the cosine angle between them or their distance relationship within the vector space. The higher the dot product, the closer the vectors in vector space, which indicates a higher degree of similarity. On the other hand, the same applies to the smaller Euclidean distance between the vectors, which also indicates higher similarity between the associated images in the archive to the query. Examples of text-based medical image retrieval systems available in the literature include FigureSearch [YC08], BioText [HDG+07], Yale Image Finder [XMK08], GoldMiner [KJT07], RADTF [DWB+10] and iMedline [GAL+11].

Usually, text-based retrieval systems are fast since the metadata can be structured, making it optimal for the search process. In contrast, their reliability is highly affected by the accuracy and availability of the metadata. For example, to determine the similarity between medical images, the first and crucial step is categorizing them based on their imaging modalities [GKK+02]. Unfortunately, modality is often not correct or present

in metadata information that can be extracted from the associated captions of medical images [Ima12]. Under such circumstances, the retrieval quality would be reduced significantly.

In other words, this dependency on annotation is the main strength of text-based retrieval systems, as it can guarantee fast results. Ironically, it is also the main drawback since it is, unfortunately, costly to obtain annotated medical images in clinical settings let alone their accuracy is not guaranteed. Annotating medical images is challenging for clinicians because it is time-consuming and tedious. In addition, it is also error-prone because it is difficult to describe the content of medical images with limited words [CLQ+19a; MSD+20; HAA+20; Mül20].

### 4.2.1 Feature Representation

For text-based medical image retrieval systems, feature representation is needed for both the query and the associated metadata for the images in the archive. As explained in Definition 4.1, the feature representation process acts as a framework ($\mathcal{F}$) that model the data representations and the queries with a feature transformation function $f(\cdot) \in \mathcal{F}$. The overall transformation process can be split into two steps: first, term selection, and second, feature vector generation [LLC+13]. Many systems employ techniques to remove or conflate common words to standard forms in the term selection step. This comprises the removal of *stop words*, which are mostly used words that usually occur with high frequency and are generally of little value in search. The list of stop words, also known as a negative dictionary, varies in magnitude from the seven terms of the original MEDLARS list (by, an, and, of, the, from, with) to the list between 250 to 500 words more widely used. Examples of the latter are the list of 250 words by van Rijsbergen, 471 words by Fox [FFY92], and the PubMed stop list [oMe07]. The conflation of terms to common forms is executed through stemming with the goal of ensuring words with plurals and common suffixes (e.g., -ing, -al, -ed, -er) are represented through their stem form [Fra92]. For example, the words cough, coughing, and cough are all indexed via their stem cough. Removal of stop word and stemming help produce a compact representation that leads to more efficient query processing [Her15].

In the feature vector generation step, a commonly used approach has been a *Term Frequency Inverse Document Frequency (tf-idf)*. tf-idf is a statistical measure used to determine the mathematical significance of words in documents [Aiz03] and hence the vectorization process is based on weights assigned to the words. *Term Frequency (tf)* is a measure of the frequency with which a term ($t$) occurs in a given document ($d$) and is assigned to each term in each document, with the following Equation 4.1:

$$tf(t,d) = f(t,d) \tag{4.1}$$

where $f(t,d)$ is frequency of term in document.

The $idf$ value is the logarithm of the ratio of the total number of documents to the number of documents in which the target term occurs. It is assigned once for each term in the collection, and it correlates inversely with the frequency of the term in the entire collection. The equation to compute the $idf$ value is:

$$idf(t) = log\frac{N}{|\{d \in D : t \in d\}|} \tag{4.2}$$

where $N$ represent the total number of documents in the corpus $N = |D|$ and $|\{d \in D : t \in d\}|$ is the number of documents where the term $t$ appears (i.e., $tf(t,d) \neq 0$). If the term is absent in the corpus, it will be a division by zero. Therefore, altering the denominator to $1 + |\{d \in D : t \in d\}|$ [TFE23] is common. Eventually, the weighting of tf-idf combines the two terms($w$) [Her15]:

$$w(t,d) = tf(t,d) * idf(t) \tag{4.3}$$

Considering the frequency of terms alone is not always practical in medical text due to the fact that crucial terms may not occur frequently in medical records or reports. Therefore, the overall term frequency may be a weak indicator of the importance of certain words in documents. Usually, the term in the medical texts determines the main meaning [LLC+13]. To address this, some studies introduced the Boolean model within tf-idf in which, instead of measuring the frequency of the term, a value of 1 is assigned while the document contains the key term. Otherwise, 0 value is assigned [LLC+13]. Another drawback of tf-idf and other approaches alike, e.g., BM25, is that they cannot capture the semantic association between words, which is highly important in determining the semantic meaning of medical texts. Luckily, the advent of deep learning has currently brought advanced text representation techniques such as Word2Vec [MCC+13], Glove [PSM14], BERT [DCL+19] e.t.c that can be applied in medical settings. In the abstract, the main idea in these deep learning models is to learn the relationships between terms and therefore infer the semantics and contextual information in medical texts hence they can significantly improve retrieval when deployed in text-based medical image retrieval systems.

## 4.3 Content-Based Retrieval

Among many things, content-based retrieval aims to complement the weakness of the text-based approaches by relying on the actual contents of medical images rather than their annotations to identify and retrieve similar cases. This, however, is not a trivial task primarily because of the semantic gap between computers' mathematical nature of interpreting images and radiologists' trained interpretation of medical images (see Section 1.2). Therefore it is of utmost importance for a content-based retrieval to utilize feature representations techniques that can abstract high-level semantic concepts from images to be close to how a radiologist sees medical images.

To illustrate the importance of encoding semantic concepts in the feature representation. Let us consider the following situation; for reference purposes, we call it *situation X*. A radiologist has a CXR image with local white opaque patches on the left lung. This is a sign of consolidation, pneumonia, or both [RZK+19] and would still be the same diagnosis even if the same patches were on the right lung. In this situation, if the representation encodes the features by linking with the position of the patches. It means, during the retrieval, only images with similar patches on the left lungs will be retrieved as similar images. Other images that still have pneumonia or consolidations would be ignored. In turn, some images that would give vital information to a radiologist for a detailed comparative analysis would be missing and hence, reducing a radiologist's ability to make an informed clinical decision.

### 4.3.1 Feature Representation

In images, feature representation methods usually aim to describe image features using pixel values by constructing feature vectors that model specific information considering low-level image signals like texture, shape and color. Said differently, low-level image signals act as input to the feature transformation function, $f(\cdot) \in \mathcal{F}$ that the feature representation methods depend on developing the framework, $\mathcal{F}$ for data, $\mathcal{D}$ and query, $\mathcal{Q}$ representations (refer Definition 4.1). Generally, image feature representation methods can be classified into two categories: handcrafted and learned features [LZM+18].

#### 4.3.1.1 Handcrafted Features

Handcrafted features are extracted from images following algorithms designed by experts [LZM+18]. In simpler words, handcrafted features are manually designed features. These features aim to characterize an image by focusing on specific issues like occlusions and variations in scale and illumination [NGB17] e.t.c. In medical imaging, texture has been a dominant signal to characterize an image compared to other signals like color

and shape. This has been the case for two reasons. First, most medical imaging modalities produce grayscale images; therefore, signals like color only add a little valuable information. Second, a texture signal represents the spatial distribution of an image's pixel values; thus, they are helpful in medical images as they can reflect the detail within an image structure [HLC12].

In addition, medical images contain an incredible amount of texture information that is useful for clinical practices compared to other signals. For example, some MRI images of tissues may not provide microscopic information that can be assessed visually. However, histological alterations in some diseases may cause changes in the texture of the MRI image that are susceptible to quantification using texture analysis. This has been successfully applied to classifying pathological tissues from the lungs, kidneys, heart, liver, breasts, prostate, thyroid, and brain [CBL+04]. In general, there exists a significant number of methods to express texture features for medical images, and some of these methods include the followings:-

**Haralick's Textures:** Haralick's textures compute the representation of image texture features based on the Gray Level Co-occurrence Matrix . This statistical method extracts information about pixels' positions with similar gray levels values in a two-dimensional array [HSD73]. Both rows and columns of the array represent a set of possible image values that can be defined by first specifying a displacement vector $d = (d_x, d_y)$ and counting all pairs of pixels separated by $d$ having gray levels $i$ and $j$. So the resulting matrix $P$ is $P(i, j) = n_i * n_j$ where $n_{i,j}$ is the number of occurrences of the pixel values $(i, j)$ lying at distance $d$ in the image [Mic04].

Using four co-occurrence matrices, Haralick [HSD73] proposed measures of fourteen textural features: Angular Second Moment, Maximal Correlation Coefficient, Difference Entropy, Sum Variance, Entropy, Correlation, Inverse Difference Moment, Contrast, Information Measures of Correlation, Difference Variance, Sum Entropy, Sum Average, and Sum of Squares Variance which describes the correlation in the intensity of pixel values spaced next to one another. It also details how the intensities in the image's pixel values in a particular position relate and occur together [DMM19].

Among all, entropy, Homogeneity, and correlation are the three most prominent Haralick's features and most responsible in expressing medical images [DMM19]. *Entropy* (4.4) measures the Homogeneity and randomness or Homogeneity in pixel distribution with respect to orientation or length. It usually increases in value as the randomness in the distribution increases. *Homogeneity* (4.5) refers to the similarity between different parts of the image, which means a homogeneous image will have a co-occurrence matrix with a combination of high and low $P(i, j)$ while a composite image will have an even spread of $P(i, j)$. *Correlation* (4.6), on the other hand, measures how a pixel is

correlated to its neighborhood, which is the measure of linear dependencies in a gray tone of the image.

$$Entropy = (-1) \sum_i \sum_j P(i,j) \ln (P(i,j)) \tag{4.4}$$

$$Homogeneity = \sum_i \sum_j \frac{1}{1+(i-j)^2} P(i,j) \tag{4.5}$$

$$Correlation = \frac{\sum_i \sum_j P(i,j) - \mu_i \mu_j}{\sigma_i \sigma_j} \tag{4.6}$$

In here, $\mu_i$ and $\mu_j$ are the mean of the $i$ and $j$ pixels, respectively while $\sigma_i$ and $\sigma_j$ are the standard deviations.

**Local Binary Patterns:** LBPs is another texture feature representation technique in which the representation is constructed by comparing each pixel with its surrounding neighborhood of pixels. Constructing the LBP textural features involves converting the image to grayscale and then for each pixel in the image, number of neighbours $p$ within a radius $r$ are selected to calculate the LBP value (see Equation 4.7). The result is a two dimensional array with the same height and width like the input image. This texture descriptor is robust in terms of grayscale variations since, by definition, it is invariant against any monotonic transformation of the grayscale [OPM02].

$$LBP = \sum_{p=0}^{p-1} f(n_i - n_c) \times 2^p \tag{4.7}$$

where $n_i$ is the $i$th neighboring pixel and $n_c$ is the central pixel. The values of the function $f(x)$ are always 1 if $x \geq 0$; otherwise, it is 0.

**Tamura Textures:** Tamura et al. [TMY78] proposed six human visual perceptual texture features built upon the results of psychological experiments. These features were regularity, directionality, coarseness,line-likeness, contrast, and roughness. According to Xiaoming et.al [XNH+18], the most expressive features for medical images are coarseness, contrast, and directionality. *Courseness* measures the size of texels, the primitive elements composing the texture. Its computational process is as follows.

Step 1: First, the mean for the brightness of all pixels over the neighborhood of size $2^k \times 2^k$ (e.g., $32 \times 32$ when $k$=5) is calculated using equation 4.8.

$$A_k(x,y) = \frac{\sum_{i=x-2^{k-1}-1}^{x+2^{k-1}-1} \sum_{j=y-2^{k-1}-1}^{y+2^{k-1}-1} g(i,j)}{2^{2k}} \tag{4.8}$$

where $(x, y)$ denotes the position of the region in the selected image, and $g(i, j)$ denotes the mean of the brightness in the $i^{th}$, $j^{th}$ points for the selected region and $k$ determines the pixel's range.

Step 2: Then, the mean intensity difference between neighborhoods that do not overlap and are on opposite sides of the point in both vertical and horizontal orientations is calculated with equation 4.9 and equation 4.10

$$E_{k,h} = \left| A_k(x + 2^{k-1}, y) - A_k(x - 2^{k-1}, y) \right| \tag{4.9}$$

$$E_{k,v} = \left| A_k(x, y + 2^{k-1}) - A_k(x, y - 2^{k-1}, y) \right| \tag{4.10}$$

where $E_{k,h}$ denotes the horizontal difference of this pixel and $E_{k,v}$ denotes the vertical difference of this pixel.

Step 3: For each pixel, we then pick the optimal size which brings the maximum output value $S_{best}$:

$$S_{best}(x,y) = 2^k \tag{4.11}$$

where $k$ maximizes $E$ in either direction, i.e.,

$$E_k = E_max = max(E_1, E_2..., E_h) \tag{4.12}$$

Step 4: Finally, we get a measure of image's coarseness $F_{crs}$ by taking the average value of $S_{best}$ for all pixels of the image (Equation 4.13)

$$F_{crs} = \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{i=1}^{n} S_{best}(i,j) \tag{4.13}$$

where $m$ and $n$ are the optimal width and height of the image, respectively. *Contrast* represents the brightness level between an image's darkest and brightest spots. Intuitively, this is a critical feature, as we explained in Chapter 2, that one clue radiologists look for when interpreting medical images is the light intensity variations, as it gives information on the tissues' absorption of radiation. The contrast of medical images is calculated as follows:

$$F_{con} = \frac{\sigma}{\alpha_4^{1/4}} \tag{4.14}$$

where $\sigma$ is the standard deviation, and $\alpha_4$ is the kurtosis of gray values of the medical image, respectively.

*Directionality:* The directionality is of crucial importance in medical images, as it represent the texture orientation of the human muscles or tissue [XNH+18].

The calculation procedure for directionality is as follows: i) The modulus of the gradient vector and the local edge direction of each pixel are calculated using Equations 4.15 and 4.16.

$$|\triangle G| = \frac{|\triangle H| + |\triangle V|}{2} \tag{4.15}$$

$$\theta = tan^{-1}\left(\frac{\triangle V}{\triangle H}\right) + \frac{\pi}{2} \tag{4.16}$$

Both Equations 4.15 and 4.16 can be obtained through convolving a $3 \times 3$ rectangular area around an image's pixel point with two $3 \times 3$ masks.

ii) Initially, divide the region 0 - $\pi$ into 16 similar parts and acquire the angle $\phi$ corresponding to the highest mode of the gradient vector in each similar interval [XNH+18]. Then, compute pixel number $n_p$ when $|\triangle G|$ in each region corresponding to the angle $\theta$ is larger than the threshold. Secondly, compute the value of gradient vectors of all pixels to build the histogram $H_d$ and discretize the range values in the histogram. Afterward, the peak position of $H_d$ can be represented by $\phi_p$. Finally, by considering the sharpness of the peak in the histogram, the overall direction of the medical image can be calculated( 4.17).

$$F_{dir} = \sum_{P}^{n_p} \sum_{\phi \in \omega_p} \left(\phi - \phi_p\right)^2 H_d(\phi) \tag{4.17}$$

where $p$ represents the peak and $\omega_p$ represent the range of the peak between each valley. Tamura texture features have been successfully applied for content-based retrieval by many studies including [XNH+18; ZFL+12; MS12; PK11; GMK11; XSC+10].

**Scale-Invariant Feature Transform:** Scale-Invariant Feature Transform (SIFT) is the basis for most commonly used handcrafted features in image retrieval tasks [Low04; LZM+18]. SIFT recognize scale-invariant key points through searching for local extrema in the difference-of-Gaussian (DoG) space, which describes each key point by a 128-dimensional gradient orientation histogram. Subsequently, all SIFT descriptors are quantized using a bag-of-words (BoW) [SZ03]. The feature vector of each image is calculated by counting the frequency of the image's generated visual words. SIFT, as a local texture feature has gained significant success in the retrieval of medical images (e.g., it was the top used feature in the 2012's ImageCLEF's medical image retrieval competition [MHK+12]) and also applied by [ZLD+15].

One advantage of the texture feature representation methods mentioned above is their applicability to different medical imaging modalities, making them useful for general-purpose medical image retrieval systems. However, since each modality presents its unique feature, some researchers have designed specific texture features [LZM+18]. For example, to encode texture information vital in representing cell/nuclei in histopathology images, Basavanhally et al. [BGA+10] tailored three graph-based features, i.e., minimum spanning tree, Voronoi diagram, and Delaunay triangulation, and to describe the structure of the lymphocytes. On the other hand, Filipczuk et al. [FFK+13] designed 25 features to represent cytological images while considering the nuclei size. The texture features were based on gray-level pixels and the distribution of nuclei in the image. These features were more distinctive compared to the general-purpose handcrafted features, and they achieved an excellent performance in detecting, retrieving, and analyzing cells and nuclei [XY16].

Apart from histopathological images, specifically designed features are widely used to represent 3D medical images, like neuronal morphology and 3D brain tumors. A good example is the work by Cai et al. [CLW+10] in which they developed PCM-based volumetric texture features to retrieve 3D neurological images. Another work is by Wan et al. [WLQ+15], who designed geometrical moments and quantitative measurements as features to represent the 3D neuron morphological images.

Despite handcrafted features having obtained many excellent results in content-

based medical image retrieval, they still have drawbacks, as explained in the following:

i. Handcrafted features require expert knowledge; however, expert knowledge does not always work well, especially when the dataset is large since there is a high chance of outliers and rare cases that are not covered by standardized rules to exist [LZM+18];

ii. Many handcrafted approaches are explicitly designed for the particular medical data, and therefore they are unexpendable to others;

iii. Handcrafted methods struggle to capture high-level semantic concepts of medical images; hence, they are limited in reducing the semantic gap.

Accordingly, there is a need for more extensible, automatic, and efficient feature representation methods for the efficient retrieval of medical images [LZM+18]. These methods should also be able to abstract high-level concepts of medical images.

#### 4.3.1.2 Learned Features

Unlike handcrafted features obtained through domain expert knowledge, learned features are obtained purely through data-driven procedures [LZM+18]. As explained in Section 1.2, it is cumbersome for a computer to infer the semantic meaning of raw input signals, like the pixel values that represent the image. However, by using deep learning, a computer can learn the mapping of raw sensory input data to output and feature representations itself [GBC16d]. These learned representations often encode high-level semantic information compared to handcrafted features. Therefore, they can reduce the semantic gap between radiologists' and computer's interpretation of medical images. On the other hand, they can solve the selectivity-invariance dilemma- which means they can be selective to the aspects of the image that are important for discriminating a respective image [LBH15] with other images, which is a crucial function for medical image retrieval systems.

Due to its success in computer vision tasks (refer Section 3.3), such as image retrieval, CNNs have been widely used to learn the representation of medical images. One approach has been to first train a CNN in a supervised learning task like image classification through which it learns the feature representations. Afterward, a model is used as a feature extractor for the retrieval task. One advantage of CNNs is that they are translational invariant, meaning they work equally well across positions, and their response shifts with the target's position. Referring to *situation X*, we introduced at the start of Section 4.3, it means a CNN would still detect pneumonia, no matter whether the local white opaque patches would be on either side of the CXR image.

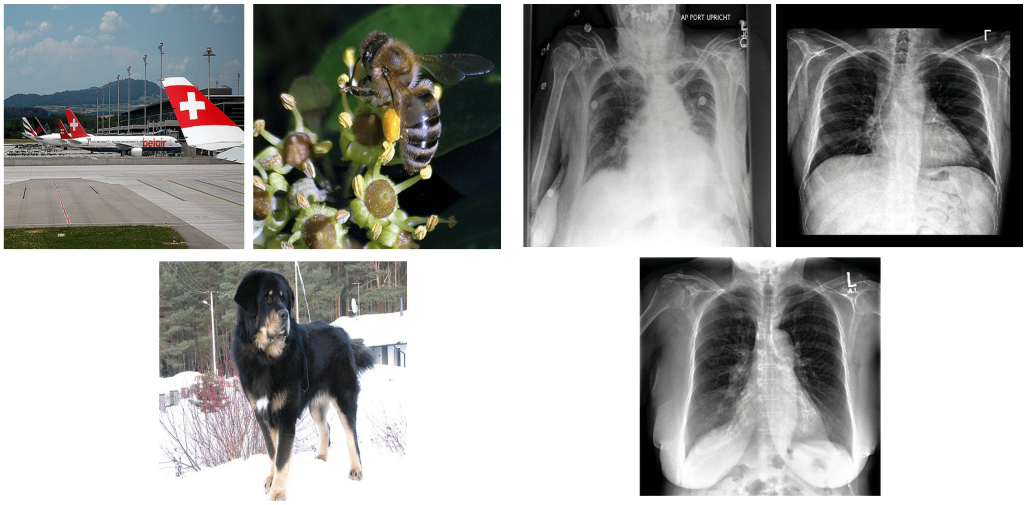


**Figure 4.1 Example of natural images (from ImageNet dataset [DDS+09]) and medical images (from CheXpert dataset [IRK+19] ).**

However, the challenge with CNNs is that they require a large amount of labeled data to train. It is cumbersome to get large amounts of data in clinical settings due to confidentiality issues, among many reasons. To deal with such a drawback, many researchers have explored different approaches to adapt CNNs for medical images. One of those approaches has been to apply transfer learning, in which a model is first pre-trained on natural images, usually the ImageNet dataset, and then used as a feature extractor or fine-tuned on the targeted medical images dataset [HL15a; AAK+19; BDW+15; TSG+16; SCN+22]. With this approach, however, one must consider that medical images are fundamentally different from natural ones. Usually, in medical imaging, there is a large bodily region of interest, and local textures' variations can identify pathologies. For example, when considering retinal fundus images, tiny red dots indicate microaneurysms and diabetic retinopathy; in CXR images, local white opaque patches are indications of consolidation and pneumonia. In contrast, this is different from natural images, where often, there is a distinctive global subject of the image like planes, dogs, bees, etc., as Figure 4.1 illustrates. Thus, there is still an open question of if ImageNet-trained models can reuse the learned features for use in medical images [RZK+19]. Another approach explored has been to train a CNN entirely from scratch in the available small medical image datasets but putting several measures in place to avoid overfitting and class imbalances issues [CLQ+19a]. Also, using a pre-trained model directly to extract features and use these features to complement handcrafted features by integrating them [LZM+18].

Even though CNNs have proved very useful in creating highly performant feature representations, their representations are usually of higher dimensions increasing computation cost for the retrieval process. To overcome this drawback, Autoencoders, which

are dimensionality reduction neural networks, can significantly help. Autoencoders present two main advantages. First, they can learn features completely unsupervised and, therefore, alleviating the problems of depending on sometimes uncertain medical image labels. Second, they can learn a lower-dimensional feature representation which can significantly reduce the computational cost required during the similarity search process to retrieve medical images. Some works that explored autoencoders includes [WKW+15]', who developed a deep feature representation method using stacked autoencoder to learn features in brain MRI images. On the other hand, Daoud et al. [DSH+19] built an autoencoder model to learn deep representations for retrieving breast cancer ultrasound images. Other works includes [Özt20] and [SUO+16].

## 4.4 Multimodal Retrieval

From the feature representation vantage point, most multimodal approaches leverage both text-based and content-based representations to inform the retrieval process. However, the main challenge in multimodal retrieval is the fusion of information from both text and image media since there is an intrinsic difference between them in expressing information. In the literature, the fusion for multimodal retrieval can roughly be categorized into feature fusion (early fusion) and retrieval fusion (late fusion). The feature level fusion integrates features from both text and image to a compound representation used for the similarity search. While in the retrieval fusion, text-based and content-based retrievals are performed separately, and then the retrieval results are combined to form a final ranked list.

Early fusion approaches have yet to be widely explored compared to retrieval fusion. This is because it is difficult to determine how textual and image information should contribute to the compound representation. An example of research in the feature fusion category includes a work by Cao et al. [CLM+11], who developed a method in which both features were represented as a multi-dimensional matrix. Finally, the feature vectors were incorporated by applying an extended Probability Latent Semantic Analysis model (pLSA). In 2021, Yu et al. [YHL+21] also developed a multimodal multitask deep learning approach for retrieving radiological images in which a model is trained to learn semantic feature representations for both texts and images and maps these representations into a common vector space. During retrieval, the representations from the common vector space are used to measure similarities among the query and the multimodal database.

In late fusion, popular techniques studied are those that perform the text-based method first and two methods at the same time [HLC12]. For example, Demner-Fushman et al. [DFAS+10] conducted three different experiments starting with text as follows:- (i)

*Text-to-Content*: textual search was first conducted, and then a mean vector of 3-5 highest ranked retrieved images which were manually selected as a query for visual search. (ii) *Text re-rank*: textual search is performed first, and then the retrieved images are re-ranked based on the scores of the visual search. (iii) *Interactive text-content*: Users manually selected relevant images from the top ten retrieved images for each query. They then selected additional query terms from the relevant images' document and used this as the input to the textual search. The additional retrieved images by the expanded query were ranked below images that were manually selected as relevant. This method performed better compared to the other two. In general, the fusion of both texts and images has proven to be effective as Zhang et.al [ZSC+17] performed experiments and found that given different methods in comparison; the text-based methods outperform the content-based retrieval. However, the fusion of text and visual content generates the best performance overall. Other studies that combined both retrieval results at the same time includes [DAED+15; LLC+13; GMK11; GAL+11; DAK10].

## 4.5 Retrieval Systems Evaluations

The ways to evaluate retrieval systems' performance can be based on either unranked or ranked results. A large number of IR studies, however, have shown that users of retrieval systems are more likely to pay attention more to top-ranked results [MC17], and therefore, rank-based measures are preferable. For radiologists, an effective system is highly accurate in the top-ranked results. Consequently, such a system requires less effort to identify relevant images for comparative analysis that supports radiologists' decision-making process. Said differently, an effective retrieval system should significantly reduce the time radiologists require to interpret medical images and improve their diagnostic accuracy [Her15; RHP+22] in the process. In this section, we briefly describe popular metrics used in the evaluation of retrieval systems:

**Precision and Recall:** Both precision and recall evaluate the relevance of the retrieved results to the information need of the user expressed through a query, $q$. Precision (equation 4.18) computes the fraction of which results are relevant in all retrieved results while recall (equation 4.21) considers, the number of relevant documents retrieved with respect to all relevant results available in the database:-

$$precision = \frac{\left|\mathbb{F}_q \cap \mathbb{R}_q\right|}{\left|\mathbb{F}_q\right|} \tag{4.18}$$

$$recall = \frac{\left|\mathbb{F}_q \cap \mathbb{R}_q\right|}{\left|\mathbb{R}_q\right|} \tag{4.19}$$

where, $\mathbb{R}_q$ is a set of relevant documents for a query $q$ in collection $\mathbb{A}$ that contains the set of all documents and $\mathbb{F}_q$ is a set of documents retrieved by a system for query $q$. When precision and recall are combined through weighted harmonic mean, we get a F-measure or F1-score which is computed as follows;

$$F1 = \frac{(1+\beta^2) \cdot precision \cdot recall}{\beta^2 \cdot precision + recall} \tag{4.20}$$

where:

$$\beta^2 = \frac{1-\alpha}{\alpha} \tag{4.21}$$

where $\alpha \in [0,1]$ and thus $\beta^2 \in [0,\infty]$. The choice of evaluating a system based on either precision, recall, or F1 score is largely determined by the types of queries the system executes and the data composition in the dataset searched through.

**Average Precision:** Average Precision (AP) is a measure that integrates precision and recall for ranked retrieval results. For one's information need, the AP is the mean of the precision scores following each relevant document retrieved ( 4.22). AP is computed as follows;

$$AP = \frac{\sum_r P@r}{R} \tag{4.22}$$

where $r$ is the rank of each relevant document, $R$ is the total number of relevant documents and $P@r$ is the precision of the top-$r$ retrieved documents. AP is considered a reasonable evaluation measure for emphasizing returning more relevant documents earlier [ZZ09].

**Mean Reciprocal Rank:** The Reciprocal Rank (RR) measure computes the reciprocal of the rank at which the first relevant document was retrieved [Cra09].The value of RR equals 1 when a relevant document was retrieved at the top position; if not, it is 0.5 if a relevant document was retrieved at the second position, and so on. Mean Reciprocal Rank (MRR) is the result of averaging the results across several queries. MRR models a scenario where the user only wishes to retrieve and hence see the relevant documents [Cra09]. Therefore, the assumption here is a user will look down the rankings

until he/she finds a relevant document, so if the document is found at rank $i$, then the precision for the viewed set is $1/ranki$, which is also the reciprocal rank measure (see equation 4.23).

$$MRR = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{rank_i} \tag{4.23}$$

where $Q$ is the size of query and $rank_i$ denotes the rank of the relevant first-ranked item in the $i$-th query.

**Discounted Cumulative Gain:** Discounted Cumulative Gain (DCG) is based on two assumptions: i) First, highly relevant documents are more important than marginally relevant documents. ii) Second, the lower the ranked position of the relevant document, the less important it is to the user, as it has a lower chance to be examined [MN19]. Therefore, DCG applies a graded relevance from examining a document as a measure of importance, or *gain*. Gain is accumulated at the top of the ranking and may be reduced or discounted as the ranks decrease. So, in general, DCG is the total gain aggregated at a specific rank position $p$.

$$DCG_p = \sum_{i=1}^{p}\frac{2^{rel_i}-1}{log_2(i+1)} \tag{4.24}$$

where $rel_i$ is the graded relevance of the retrieved document at position $i$.

## 4.6 Summary

This chapter briefly reviewed conventional approaches for medical image retrieval systems: text-based, content-based, and multimodal. We further explained the feature representation techniques for each approach and the retrieval evaluation metrics for retrieval systems in general. Feature representation is the first and critical step in any medical image retrieval system. In other words, a good feature representation is a prerequisite for achieving optimal performance [LZM+18]. The second step is the similarity search. In the next chapter, we give a detailed overview of similarities in medical images.

# 5

# Similarity in Medical Images

The performance of any medical image retrieval system depends on one key idea; there is a similarity between the information need expressed by the radiologist through the query and relevant images available in the archive. As described in the previous chapter, feature representation is one step toward this goal. So any feature vector inherently defines some notion of relatedness between medical images [MC17]. The second step is similarity search which involves determining the degree of similarity between different cases of medical images. This chapter briefly reviews the notion of similarity for retrieval systems in general and dives into similarities in medical images from mathematical and clinical perspectives. We then highlight the importance of considering both perspectives to retrieve medical images accurately.

## 5.1 The Notion of Similarity

Similarity search has become crucial in an age of large information repositories where the objects contained do not possess any specific order, for example, an extensive collection of sounds, images, and other sophisticated digital objects [TFE22f]. In general, similarity search is a range of mechanisms that share the concept of searching (typically, extensive) in spaces of objects where the only comparator available is the similarity between any object pairs. In other words, a similarity search is a comparative analysis between a pair of objects, and there must be a measure of similarity between these objects.

The notion of similarity in nature is very context-dependent, meaning that what makes a pair of objects similar primarily changes according to the information the user needs in that particular scenario. In clinical settings, in general, understanding the context is very important. It can make a difference between life and death due to the care administered to the patient as a result of inferring the proper context. To retrieve medical

images accurately, the need to consider the appropriate context between different cases of patients and what makes some cases similar is of utmost importance. This requires consideration from both mathematical and clinical perspectives.

### 5.1.1 The Mathematical Perspective

From the mathematical perspective, we can infer the similarity between a pair of objects through a *metric space*, a concept in which a non-empty set exists together with a metric on the set [TFE22c]. Referring to the Definition 4.1, in a *metric space* is where the comparison function ($\delta(\cdot,\cdot)$) is executed.

#### 5.1.1.1 Metric

A metric is a function that quantifies the "distance" between any pair of elements in the set, also known as points. A metric function $d(x, y)$ need to satisfy the following properties for all $x$, $y$, $z$ which are members of the set:-

- *Symmetry:* $d(x, y) = d(y, x)$, the distance from $x$ to $y$ is the same as the distance from $y$ to $x$.
- *Non-negativity:* $d(x, y) \geq 0$, the distance between two distinct points is positive.
- *Identity of Indiscernible:* $d(x, y) = 0 \iff x = y$, the distance from $x$ to $y$ is zero if and only if $x$ and $y$ are the same point.
- *Triangle Inequality:* $d(x, y) \leq d(x, z) + d(z, y)$, the distance from $x$ to $y$ is less than or equal to the distance from $x$ to $y$ via any third point $z$.

In broad strokes, we can categorize metrics into two groups; pre-defined and learned metrics [CL17]. Pre-defined metrics assume that the points in the metric space are already perfect to describe the similarity/dissimilarity between data. In other words, in the retrieval model tuple as explained in Definition 4.1, the data, $\mathcal{D}$ and query, $\mathcal{Q}$ representations exist already; therefore, we can directly compute the distance between these data points as the similarity measure without much knowledge of the data—the less the distance, the high the similarity. Example of such metrics includes Minkowski distance functions. The Minkowski distance $L_m$ of order $p$ between two points $p_1 = (x_1, x_2, x_3..., x_n)$ and $p_2 = (y_1, y_2, y_3..., y_n)$ in normed vector space is defined as:

$$L_m = \left( \sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}} \tag{5.1}$$

The value of $p$ is often based on experimentation since its effectiveness in measuring similarity is based on the context, or use case applied. Therefore determining the optimal value of $p$ is critical in obtaining the correct results. When we set $p = 1$, the Minkowski distance becomes a Manhattan distance ($L_1$) which is the distance between two points measured along axes at right angles.

$$L_1 = \sum_{i=1}^{n} |x_i - y_i| \tag{5.2}$$

When we set $p = 2$, the Minkowski distance is the same as the Euclidean distance ($L_2$), which is the straight line distance between two points.

$$L_2 = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{5.3}$$

In summary, Minkowski distance is the generalized distance between two points, while Manhattan and Euclidean distances are special cases of Minkowski distance when $p = 1$ and 2, respectively. As explained above, setting value $p$ is critical in obtaining the preferred results. The illustration of this is seen in Figure 5.1 that shows unit circles (the level set of the distance function in which each point is at the unit distance from the center) with different values of $p$ [TFE22d].

Even though the Minkowski distances, especially $L_1$ and $L_2$, are widely used in the literature, there is still a shortfall in relying only on pre-defined metrics because the retrieval performance will ultimately depend only on the effectiveness of the priors, the feature representations. Learned metrics that jointly learn the feature representations and a distance metric based on the knowledge of data can significantly address this shortfall, as explained in the following section.

#### 5.1.1.2 Metric Learning

The primary goal of a metric learning approach is to learn a new metric that reduces the distances between objects of the same class and increases the distances between the objects from different classes [DLF+17]. Said differently, metric learning aims to bring similar objects closer while dissimilar objects further apart (see Figure 5.2c). In the context of the Definition 4.1, metric learning is a technique to develop the framework, $\mathcal{F}$ to model both data, $\mathcal{D}$ and query, $\mathcal{Q}$ representations end-to-end with the help of the comparison function, $\delta(\cdot, \cdot)$.

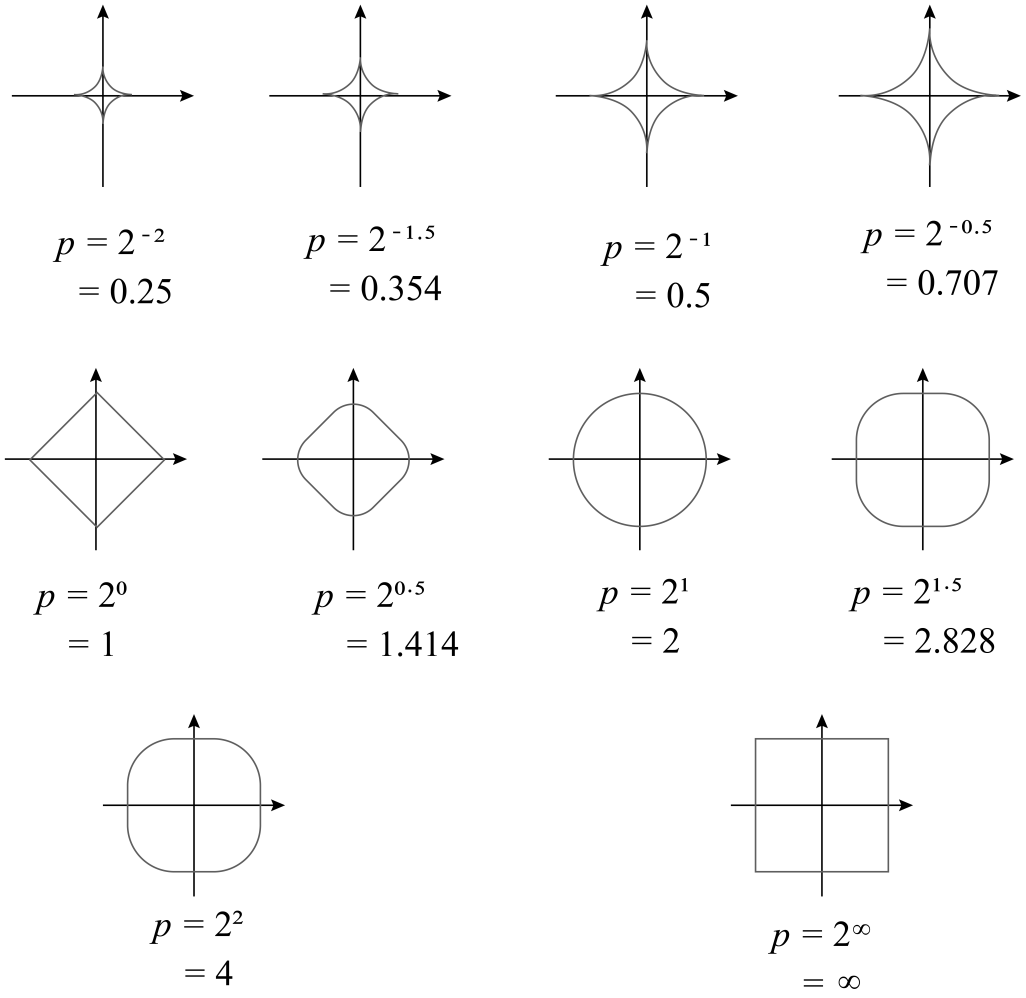


**Figure 5.1 Unit circles with various values of $p$ (Minkowski distances) [TFE22d].**

Before deep learning became prominent, the most notable example of a learned metric was the Mahalanobis distance, defined as follows; let $X = [x_1, x_2, x_3..., x_n] \in \mathbb{R}^{d*n}$ be the training samples, where $x_i \in \mathbb{R}^d$ is $i$th training example, and $n$ is the total number of training samples. The distance between $x_i$ and $x_j$ ($d_m(x_i, x_j)$) is calculated as:

$$d_m(x_i, y_j) = \sqrt{\left(x_i - y_j\right)^T M \left(x_i - y_j\right)} \tag{5.4}$$

As a distance metric, $d_m(x_i, x_j)$ must have all metric space properties, including symmetry, non-negativity, the identity of indiscernible, and the triangle inequality. On the other hand, $M$ is a matrix estimated from the data and needs to be symmetric and positive semi-definite, which means all of its determinants or eigenvalues must be positive or zero. By spectral theorem [TFE22e], $M$ can be decomposed to:

$$M = W^T W \tag{5.5}$$

where $W$ is an orthogonal matrix composed of eigenvectors of $M$, which gives us the equivalent definition of $d_m(x_i, x_j)$ to:

$$d_m(x_i, y_j) = \sqrt{\left(x_i - x_j\right)^T W^T W \left(x_i - x_j\right)} = \parallel Wx_i - Wx_j \parallel_2 \tag{5.6}$$

where $||\cdot||$ is euclidean norm and $W$ has a linear transformation property. Due to this property, the Euclidean distance in the transformed space is similar to the Mahalanobis distance in the original space for two objects.

The Mahalanobis distance metric captures the data structure and their relationships underneath, providing a new data representation in the transformed space with adequate discrimination power among classes of similar objects. However, the major limitation of this Mahalanobis distance is its reliance on the linear transformation of data. In the real world, the relationship among data is usually not linear; therefore, Mahalanobis needs to capture the true nature of relationships among data. As an alternative approach to this limitation, researchers have turned to Deep Metric Learning. Deep Metric Learning can explore nonlinearity among data, transforming the data into a non-linear space.

#### 5.1.1.3 Deep Metric Learning

Deep Metric Learning (DML) relies on deep neural networks to optimize the feature representation of the input data conditioned based on the similarity measure applied [CL17]. Unlike Mahalanobis' method, the neural networks in DML use activation functions that have a nonlinear structure, which helps to learn the nonlinear relationship among data, hence transforming the data into a nonlinear space. Consequently, DML gives a new feature representation with a more meaningful and discrimination power capable of distinguishing even subtle dissimilarity between samples (see the illustration in Figure 5.2) [KB19].

There are many ways in which we can leverage DML. One way is through a classification task in which one can achieve similarity learning simply by training a model to solve the classification problem [Agr21]. This model will then be used as a feature extractor, making the feature representations upon which the distances are computed. The assumption here is that objects from the same classes are automatically expected to have smaller distances than objects from different classes.

Another approach is to train a DML model to learn the similarity between objects end-to-end. Like any other deep learning algorithm, to build such an algorithm, four

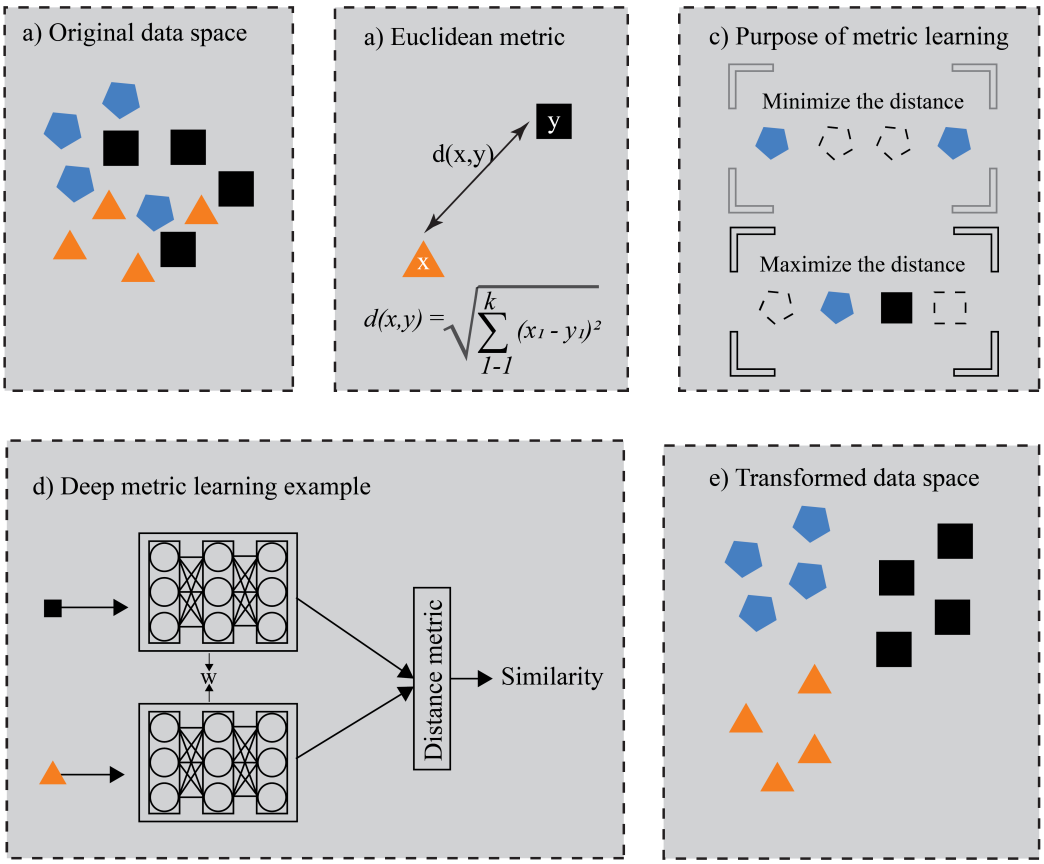


**Figure 5.2 Deep Metric Learning [KB19].**

components are required, i) a dataset, ii) a model, iii) a loss function, and iv) an optimization algorithm [GBC16e]. The choice of the loss function is mainly the most critical one. Traditionally, the widely used loss functions for DML in the literature have been the contrastive loss and the triplet loss [Agr21]. An excellent example of this is a work by Chung et al. [CW17], who proposed a deep Siamese Network (see Figure 5.3) to learn representations of Diabetic Retinopathy Fundus images for content-based retrieval.

This network had multiple symmetric subnetworks tying the same parameters and weights that update mirrorly and conjointly at the top by an energy function. The model used Rectified Linear Unit (ReLU) nonlinearity as the activation function for all layers. The similarity between images was computed using Euclidean distance, and the loss function was defined through the computation of the contrastive loss as follows:-

$$L(W, I_1, I_2) = \mathbf{1}(L = 0)\frac{1}{2}D^2 + \mathbf{1}(L = 1)\frac{1}{2}[max(0, margin - D)]^2 \tag{5.7}$$

where $I_1$ and $I_2$ are a pair of retina fundus images loaded into each of two identical

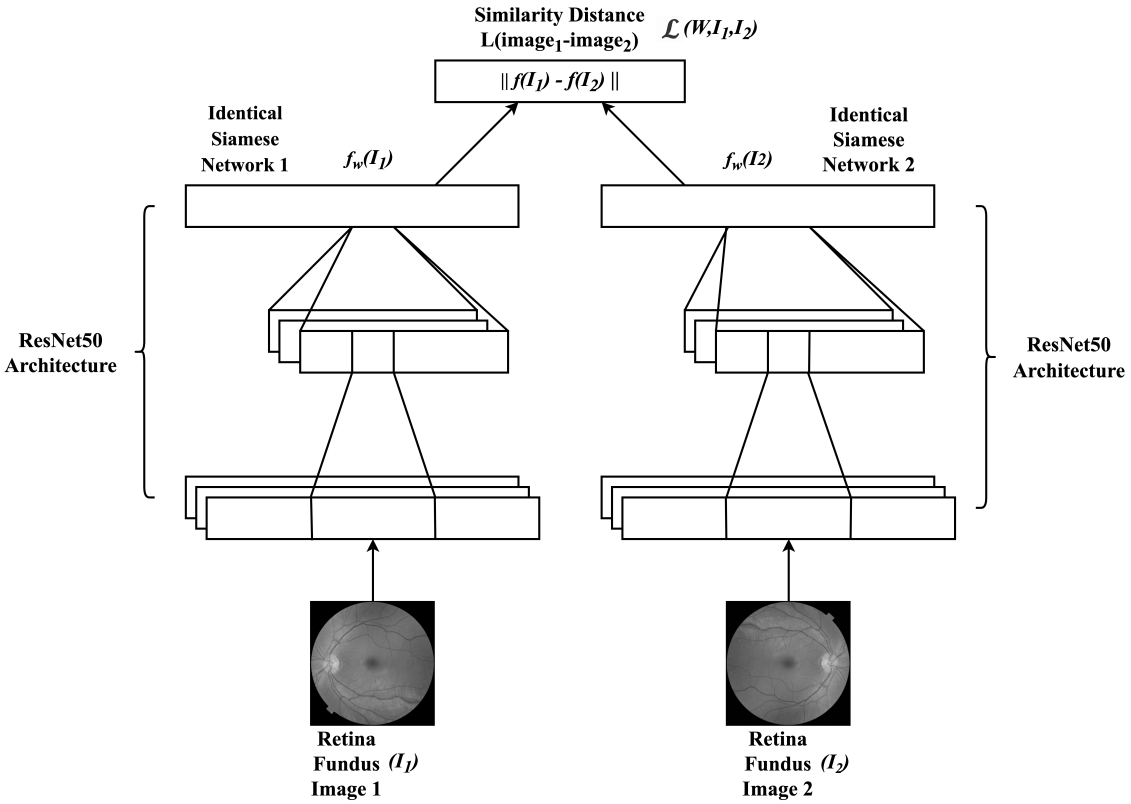


**Figure 5.3 Proposed Siamese Network by [CW17].**

networks. $1(\cdot)$ is an indicator function showing whether two images have similar labels, where $L = 0$ denotes the images have similar labels and $L = 1$ denotes otherwise. $W$ is the shared parameter that a neural network learns. $f(I_1)$ and $f(I_2)$ are the latent representation vectors of input $I_1$ and $I_2$, respectively. $D$ is the Euclidean distance between $f(I_1)$ and $f(I_2)$, which is $||f(I_1) - f(I_2)||_2$.

The limitation of contrastive and triplet losses is that they rely on the assumption that data can belong to distinct classes. Unfortunately, this is not always the case with medical images. Most medical images can have multiple diseases belonging to multiple classes. Therefore, for a DML model to perform well in medical images, one must design an appropriate loss function considering their multi-similarity nature.

### 5.1.2 The Clinical Perspective

As explained in the motivation of this thesis (see Section 1.1), diagnostic radiology is perceptual and subjective. This means radiologists' experiences are critical to their ability to reach a definitive diagnosis. Expert radiologists not only can perceive abnormalities that are hard for a non-expert to spot, but also they understand better what to attend to and what to ignore [GP19] when interpreting medical images. On the other hand,

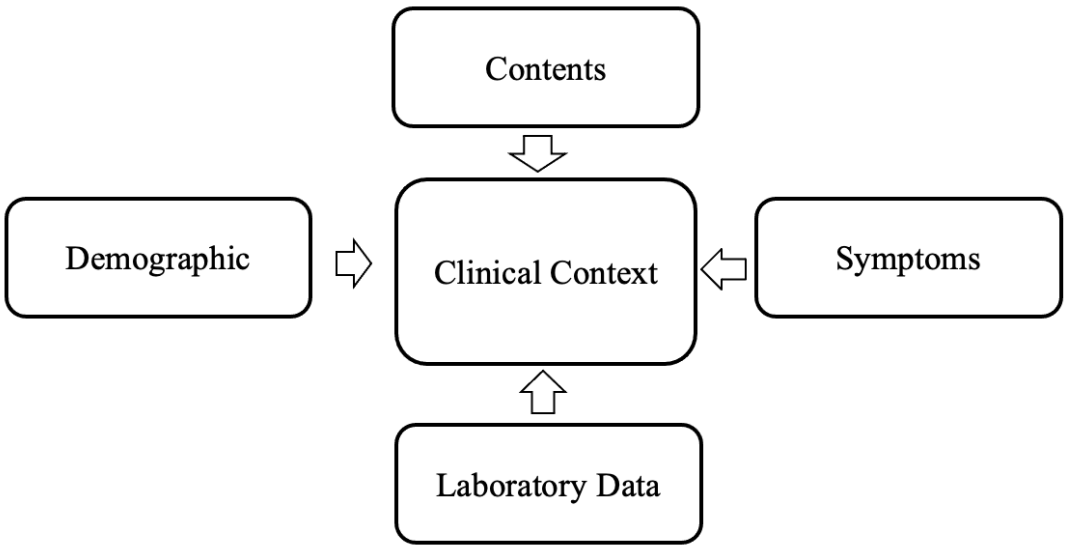


**Figure 5.4 An example of information radiologists uses to identify clinical context when diagnosing a patient. A medical image retrieval system must leverage such information to identify similarities between different medical image cases accurately.**

an expert radiologist would know the right way to consider additional patient information like demographics, symptoms, laboratory results, e.t.c., to better infer the clinical context. The importance of this additional information can be seen in the following examples—first, the age of the patients. This demographic attribute is an essential factor because it contributes to phenotypic changes in health and disease and, therefore, can affect the course and progression of a disease. It is also vital in determining the correct course of treatment [GCR13].

Another example of additional information is the results of laboratory tests. To illustrate the importance of this information, let us consider the situation when interpreting CXR images. In CXR images, similar features that indicate Pneumonia diagnosis would be accurate in one person with fever and elevated white blood cell count. However, for another patient without those supporting clinical characteristics and laboratory values, similar imaging findings may instead indicate other etiologies such as pulmonary edema, atelectasis, or lung cancer [HPS+20].

Both of these examples mean that identifying the similarity between medical image cases needs to adapt to the ways of interpreting medical images practiced by radiologists, which is more than just a correct feature representation and a proper metric function alone. Apart from image contents, the retrieval system must also incorporate other information like patients' demographics, symptoms, laboratory tests, and others when available and figure an optimal way to leverage such information to accurately identify

similarities between medical image cases (see Figure 5.4).

## 5.2 Summary

This chapter briefly overviews the second most crucial step in any medical image retrieval system: similarity search. We revisited the notion of similarity in retrieval systems and highlighted the importance of considering both mathematical and clinical perspectives to identify similarities between medical images accurately. In the next part of this thesis, we present three studies where each one highlight a retrieval method to accurately identify and retrieve similar cases based on the information need expressed by the radiologist's query.

Part III

# Retrieval Approaches

# 6

# Content-Based Retrieval

Since we ought to make a general-purpose medical image system, it must be able to handle various information needs of radiologists. As we explained in Chapter 4 and Chapter 5, to develop a medical image retrieval system, the very first step is getting the suitable method for feature representations before similarity search, which is the second step. In this chapter, we present this thesis's first study that looks at the retrieval of medical images based on their contents. Here a radiologist expresses information need by submitting a sample image (query by example), and the system computes the similarities of medical images based on their contents. This study mainly aims to find an effective feature representation method that can identify similarities between medical images by considering their semantics and modalities. We present the experiment results on different representation techniques based on handcrafted methods (mainly texture features) and deep learning (deep features). Based on these results, we propose an effective feature representation approach and deep learning architectures for learning and extracting medical image features.

## 6.1 Introduction

Radiologists need to understand how different medical imaging modalities work, what they measure, and how to interpret images produced by such modalities. Such skills can guarantee the accuracy of their diagnoses and influence how they identify similarities between images of the same or different modalities. As we explained in the Section, 2.2, medical imaging modality refers to the technique and process used to visualize a particular part of the body, organs, or tissues for diagnostic purposes. Currently the conventional modalities include X-rays, CT, MRI, US, and PET (refer Chapter 2). Depending on the modality used to diagnose a patient, the same disease can manifest different visual patterns (see Figure 6.1). This is because each imaging modality maps specific physical

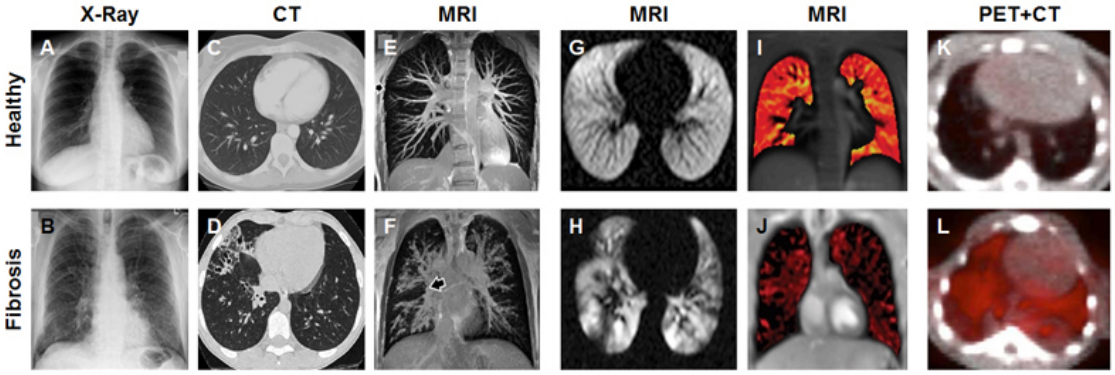


**Figure 6.1 Manifestation of Lung Fibrosis in different imaging modalities [BDE+17].**

parameters. For example, X-rays measure how different tissues absorb different amounts of radiation; therefore, the end visualization makes bones appear white since the calcium within absorbs X-ray radiation the most. Fat and other soft tissues soak in less radiation and thus appear grey. In contrast, the air absorbs the least, and therefore, lungs appear black. MRI, on the other hand, maps the energy released by protons in human body water, while US maps ultrasound backscattering.

Figure 6.1 illustrates an excellent example of how the same disease manifests different visual patterns depending on the modality used to diagnose a patient.

Here, anatomical X-ray films of the patient with pulmonary fibrosis appear dark with non-dense healthy lung tissue (A) while (B) shows a decreased lung volume and reticular opacification [BDE+17]. On the other hand, CT scans depict normal respiratory bronchioles and a lack of air spaces in healthy lungs (C); in the case of fibrosis, there are thickened and dilated respiratory bronchioles along with large cystic air spaces (D). 3D pulmonary MR angiography of a healthy lung shows well-defined vessels (E), while a fibrotic lung is characterized by enlarged, inflamed, and undefined vessels (F). The application of a hyperpolarized $^{129}Xe$ MRI contrast agent made the visualization of homogeneous ventilation in a healthy volunteer (G) possible while, in a patient with cystic fibrosis, $^{129}Xe$-enhanced MRI shows distinctive inhomogeneities in the ventilation pattern. Self-gated non-contrast enhanced functional lung MRI allowed the checking of ventilation patterns in a healthy person(I) and cystic fibrosis patient (J) without requiring a contrast agent. In healthy lungs, there was no visible accumulation of the collagen-specific PET probe $^{68}Ga - CBP8$ probe (D). In contrast, in mice that have bleomycin-induced pulmonary fibrosis, the absorption of 68Ga-CBP8 was clear [BDE+17].

While Figure 6.1 shows the manifestation of fibrosis by different modalities on different patients, the same situation can still appear in the manifestation of a particular disease to a single patient. Figure 6.2 illustrates such a phenomenon in a contrasting ex-

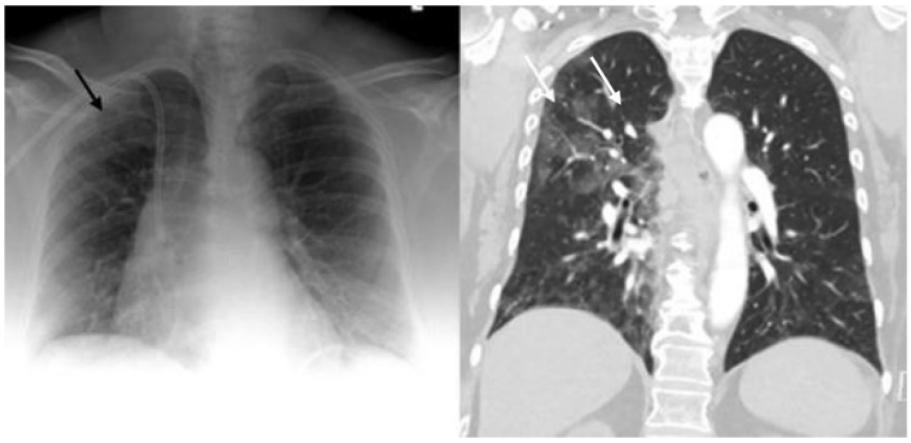

**Figure 6.2 A comparison in manifestation of COVID-19 (shown by arrows) between a CXR (left) and CT(right) images [JCB+20].**

ample between a CT and a CXR images of the same patient with COVID-19. Here, the CT image (right) shows vague hazy densities in the right upper lobe (white arrow) that corresponds to ground glass opacities (black arrow) detected in CXR (left) [JCB+20].

Both phenomenon in Figure 6.1 and 6.2 highlights one important thing. For a radiologist to do a proper comparative analysis with the image he/she wants to diagnose, it is crucial to compare it with images from the same modalities as they would present similar visual clues rather than images from different modalities. From the retrieval viewpoint, an effective retrieval system must be able to identify and retrieve images of the same modality and semantics as the query image that a radiologist wants to diagnose. This is critical, especially when the collection to be searched through contains various kinds of medical images.

A text-based medical image retrieval system is not a good solution since depending on image metadata to identify modality is not always feasible. This is because there are always changes in medical imaging protocols which increase the engineering cost for retrieval systems to adapt, let alone to different annotations procedures practiced by different healthcare institutions. To alleviate these drawbacks, content-based medical image retrieval systems have been a defacto approach since they depend solely on visual features rather than metadata to retrieve similar images. However, the challenge with this approach is that there is yet to be an optimal feature representation method to effectively distinguish medical images based on their modalities, let alone their semantics.

## 6.2 Methodology

### 6.2.1 Problem Formulation

The main question this study aims to answer is; given a collection containing medical images of various modalities and semantics, what feature representation method guarantees a good performance across them? In other words, what feature representation method is good enough to distinguish medical image modalities and semantics through their contents? To answer this question, we conducted experiments to analyze the performance of two feature representations methods i) handcrafted features (texture features) and ii) learned features (deep features) in retrieving images in a dataset containing a mixture of various medical images.

### 6.2.2 Dataset

We used the 2013 ImageCLEF medical dataset [HKD+13]. It contains different modalities of medical images, including X-rays, MRI, CT, PET, PET-CT (hardware combination of PET and CT modalities), and US. The dataset also contains other biomedical images like drawings and illustrations. There are 306,538 medical images, both of which are extracted from 75,000 articles from the biomedical open-access literature. There are 35 query topics, each containing 1-7 sample images. Each query has 1,000 ground truth image results that are supposed to be retrieved. Since our study focuses entirely on diagnostic radiology, we only selected queries from diagnostic images.

### 6.2.3 Feature Representation

#### 6.2.3.1 Texture Features

Texture features are handcrafted features, meaning their design is based on a model developed by a domain expert. In medical imaging, texture has been a dominant signal to characterize an image compared to other signals like color and shape. This has been the case for two reasons; first, most medical imaging modalities produce grayscale images; therefore, signals like color only add a little valuable information. Second, a texture signal represents the spatial distribution of the image's pixel values; thus, they are helpful in medical images since they can reflect the detail within an image structure. Various models exist to express texture features, but fewer can work across different modalities of medical images. In this study, we have experimented with Haralick and LBPs texture features. We examined the former because compared to other texture features representations, Haralick textures offer stability due to their applicability to different

kinds of images [HSD73], which has been an appealing reason to many researchers to apply them in medical image applications [DMM19; LLC+13; PK11; SCE+10]. On the other hand, we examined LBPs texture features because they have been compelling and concise texture features representations, with the capability to compete alongside state-of-the-art complex learning algorithms and quantify important textures in medical imaging [BKS+17; NLB10; AKG+15; RBH14; BTZ+17; SSG+19].To create Haralick texture feature representation, we computed all 14 Haralick's features, and for LBPs, we used a radius of size three with 24 neighbors to express the texture features. For more details about Haralick and LBPs texture features, refer to Section 4.3.1.

#### 6.2.3.2 Deep Features

As we explained in Chapter 3, a deep neural network architecture designed for the image classification task is trained on a set of images to learn features, like contour detectors and edges from earlier layers. Deeper layers usually learn to create feature filters for more complicated patterns of the inputs, like shapes, textures, or variations of features processed at early layers. These features learned in deeper layers are called *deep features*.

Deep features can integrate low, mid, and high-level features and therefore provide an abstract representation of images [ZF14; BCV13]. This combination allows them to learn semantic concepts that are sometimes impossible to capture by handcrafted features. Li et al. [LZM+18] pointed out that one way to extract deep features in medical images is through pre-trained CNNs, trained in natural images. This practice is preferred since it helps mitigate the need for training CNNs from scratch, as that requires a dataset with a vast amount of annotated medical images, which is usually tricky to obtain. Several state-of-the-art works like [HL15b; VWL+19; SPC+20] have already successfully applied this approach.

Nevertheless, one must be careful when adopting this approach, as medical images fundamentally differ from natural images (refer to Section 4.3.1.2). For example, local textures (always unclear) are essential for detecting pathologies in many medical images. In contrast, in natural images, there is often a clear global category (e.g., cat, dog, bird) [RZK+19] (see Figure 4.1). On the other hand, things like scale might not be crucial in natural images; a bird is still a bird no matter its size in the image. In medical imaging, however, pixel spacing has a known physical correspondence, and the size matters for the diagnosis, for example, the size of tumors, cell nuclei, or lesions, [GLM+21]. All these phenomena raise the question of how deep features learned in natural images are helpful in medical images. To consider that, in this study, we experimented with different CNN architectures when they are pre-trained (when they have weights learned from natural images (ImageNet weights)) and when the same architectures have random weights.

**Table 6.1 Feature Vectors**

| CNN Architecture | Dimensions |
|---|---|
| VGG16 | 512 |
| MobileNetV2 | 1280 |
| InceptionResNetV2 | 1536 |
| DenseNet201 | 1920 |
| ResNet50 | 2048 |
| Xception | 2048 |
| NASNetLarge | 4032 |

We selected the following CNN architectures from ImageNet; VGG16 [SZ14], InceptionResNetV2 [SIV+17], ResNet50 [HZR+16], DenseNet201 [HLM+17], Xception [Cho17], MobileNetV2 [SHZ+18], and NASNetLarge [ZVS+18]. The ImageNet Large Scale Visual Recognition Challenge, shortly known as ILSVRC or simply ImageNet, is a yearly challenge that evaluates algorithms for large-scale image classification and object detection tasks. The challenge allows researchers to check the progress in detection within a wider variety of objects. It also allows measuring computer vision's progress for large-scale image indexing for retrieval as well [DDS+09]. In each selected architecture, we extracted deep features using a global max pooling layer that we added to the last convolutional layer. This creates a high-dimensional representation of images as shown by their feature vectors in Table 6.1.

"The intuition of applying max pooling and not average pooling was based on its ability to select brighter pixels from the image and thus to identify the sharp features [MS20]. Given array $A$ and $B$ with elements $[1, 1, 0, 2]$ and $[1, 1, 1, 1]$ respectively, a global max pooling would choose the values 2 from $A$ and 1 from $B$ while global average pool would choose 1 from both arrays. In medical images where a single pixel can make a big difference, global max pooling seems a better choice as it can better differentiate between arrays $A$ and $B$ while average pooling would rather blur the distinctions".

## 6.3 Retrieval Performance

After creating the feature representations, we evaluated the similarity between images by computing the $L_2$ distance as a similarity metric and precision@k (where $k = 10, 30$) as the retrieval performance metric. Table 6.2 and Figures 6.3 and 6.4 shows the retrieval results.

As Table 6.2 shows, for deep features extracted by CNN architectures, we found that there is no significant difference in the retrieval performance regardless of the CNN has

**Table 6.2 Retrieval Performance for Different Medical Imaging Modalities**

| Features | X-rays | | CT | | MRI | | PET | | PET-CT | | US | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | p@10 | p@30 | p@10 | p@30 | p@10 | p@30 | p@10 | p@30 | p@10 | p@30 | p@10 | p@30 |
| *[ImageNet Weights]* | | | | | | | | | | | | |
| **ResNet50** | **0.43** | 0.331 | **0.412** | **0.336** | 0.277 | 0.212 | **0.425** | **0.208** | 0.213 | 0.142 | 0.212 | 0.167 |
| VGG16 | 0.369 | 0.269 | 0.292 | 0.276 | **0.287** | 0.213 | 0.275 | 0.15 | 0.238 | **0.192** | 0.238 | 0.158 |
| Xception | 0.323 | 0.231 | 0.354 | 0.278 | 0.204 | 0.168 | 0.175 | 0.1 | 0.162 | 0.092 | 0.162 | 0.121 |
| NASNetLarge | 0.415 | 0.31 | 0.231 | 0.191 | 0.154 | 0.127 | 0.25 | 0.192 | 0.112 | 0.025 | 0.088 | 0.071 |
| MobileNetV2 | 0.377 | **0.367** | 0.3 | 0.273 | 0.273 | 0.224 | 0.325 | 0.208 | 0.2 | 0.092 | 0.2 | 0.188 |
| DenseNet201 | 0.385 | 0.318 | 0.404 | 0.321 | 0.281 | **0.232** | 0.3 | 0.167 | 0.25 | 0.092 | 0.25 | **0.196** |
| InceptionResNetV2 | 0.315 | 0.218 | 0.231 | 0.208 | 0.153 | 0.143 | 0.175 | 0.092 | **0.26** | 0.125 | **0.263** | 0.137 |
| [*Random Weights]* | | | | | | | | | | | | |
| ResNet50 | 0.231 | 0.177 | 0.219 | 0.15 | 0.162 | 0.131 | 0.15 | 0.133 | 0.0 | 0.042 | 0.075 | 0.125 |
| VGG16 | 0.208 | 0.197 | 0.215 | 0.171 | 0.2 | 0.155 | 0.075 | 0.1 | 0.075 | 0.067 | 0.175 | 0.158 |
| Xception | 0.246 | 0.231 | 0.285 | 0.242 | 0.235 | 0.195 | 0.025 | 0.058 | 0.2 | 0.217 | 0.175 | 0.183 |
| NASNetLarge | 0.231 | 0.205 | 0.262 | 0.187 | 0.223 | 0.196 | 0.1 | 0.083 | 0.3 | 0.183 | 0.187 | 0.167 |
| MobileNetV2 | **0.323** | **0.251** | 0.242 | 0.205 | 0.235 | 0.209 | 0.1 | 0.108 | 0.224 | 0.142 | 0.325 | 0.254 |
| **DenseNet201** | 0.277 | 0.244 | **0.362** | **0.268** | **0.342** | **0.233** | **0.225** | **0.125** | **0.3** | **0.242** | **0.361** | **0.279** |
| InceptionResNetV2 | 0.285 | 0.223 | 0.158 | 0.147 | 0.146 | 0.165 | 0.05 | 0.042 | 0.075 | 0.042 | 0.175 | 0.163 |
| *[Texture Features]* | | | | | | | | | | | | |
| Haralick | **0.108** | **0.085** | 0.046 | 0.043 | 0.042 | 0.057 | **0.1** | **0.067** | **0.1** | **0.083** | **0.088** | 0.071 |
| LBP | **0.108** | 0.059 | **0.092** | **0.068** | **0.125** | **0.094** | 0.025 | 0.033 | 0 | 0.333 | 0.038 | **0.333** |

ImageNet or random weights. However, it is remarkable that deep features extracted by these architectures performed better (even if trained in natural images) compared to texture features (see Figure 6.3 and Figure 6.4). These results confirm our hypothesis that deep features provide an optimal feature representation method for medical images compared to handcrafted features, especially texture features.

Architecture-wise, deep features extracted by ResNet50 and DenseNet201 CNNs maintained a good performance across different modalities compared to features by other architectures. This result suggests that ResNets and DenseNets are effective architectures for learning and extracting medical imaging features. Therefore, we can rely on these architectures for different medical imaging tasks, whether as feature extractors in general-purpose medical image retrieval systems or as the architecture of choice for learning medical imaging features for other tasks. Our intuition as to why ResNets and DenseNets performed better comes from their fundamental architectural design. Both of them rely on using bypassing paths as the key factor to ease the training of deep neural networks but also allow feature reuse [HLM+17]. ResNets are built by stacking residual blocks, in which pure identity mappings are applied as bypassing paths (see Figure 3.7) [HZR+16]. On the other hand, DenseNets naturally integrate the properties of deep supervision, identity mappings, and diversified depth to enable layers' access to feature-maps from all preceding layers (see Figure 3.10) [HLM+17]. We hypothesize that feature reuse is critical for medical images as most are grayscale images with the same anatomical structure (e.g., chest). Therefore, they might not have as rich features as natural images. This means the deep neural network architectures like ResNets and designed to ensure feature reuse are suitable for learning medical imaging features.

Figures 6.6, 6.7, 6.9 and 6.8 shows a comparison of the top five retrieval results for ResNet50 (with ImageNet weights), DenseNet201 (with random weights), Haralick and textures for the same query, a knee arthroplasty X-ray image shown in Figure 6.5. Here, ResNet50 deep features have consistently retrieved semantically similar images of the same modality. In contrast, the others came with mixed results containing random images, sometimes unrelated to the query image. A good example is DenseNet201 and ' fifth results that were both illustration objects rather than real medical images. As explained before, both images were contained in ImageCLEF extracted images from biomedical open-access literature.

## 6.4 Discussion

This idea of using visual contents to check the similarity of medical images based on their semantics and modalities was partially inspired by a study by [DH15], which proposed

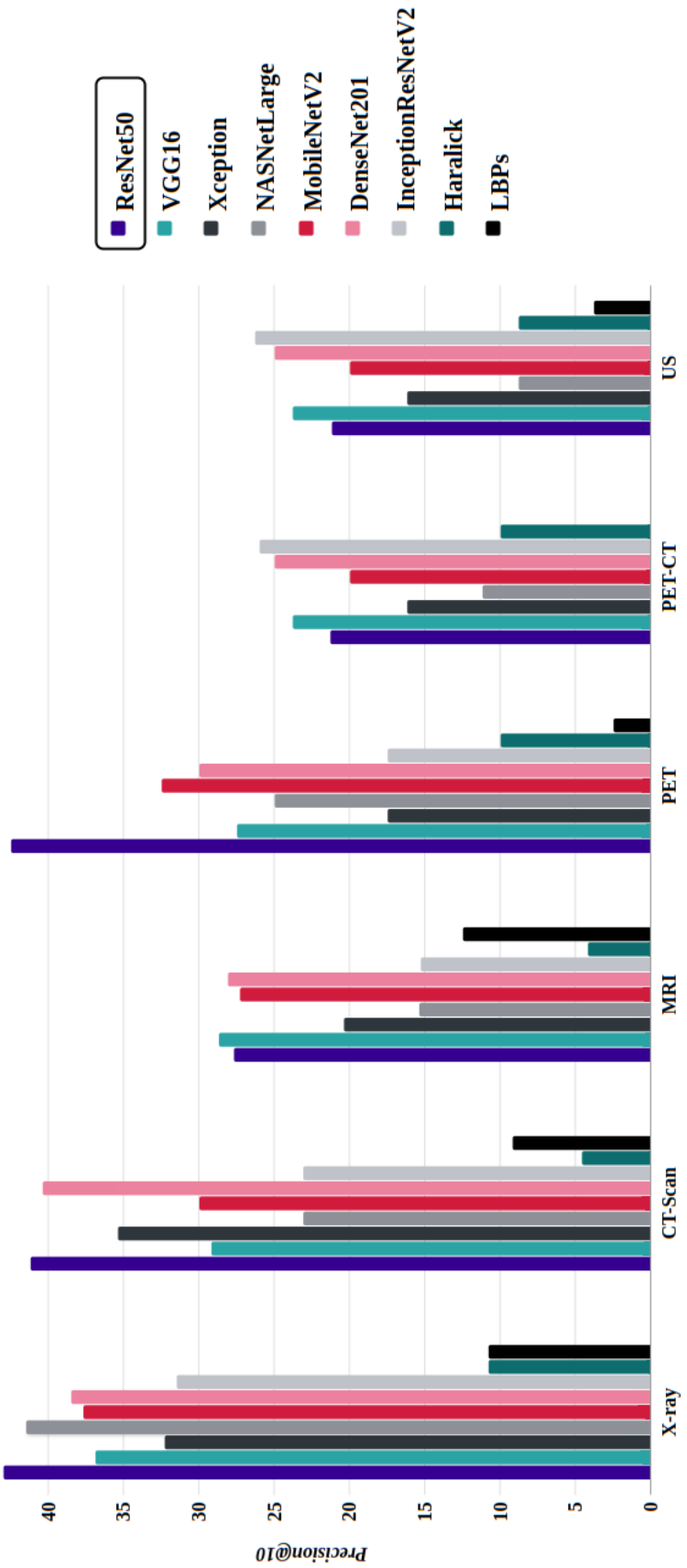


**Figure 6.3** Retrieval Performance: deep features (with ImageNet weights) vs texture features.

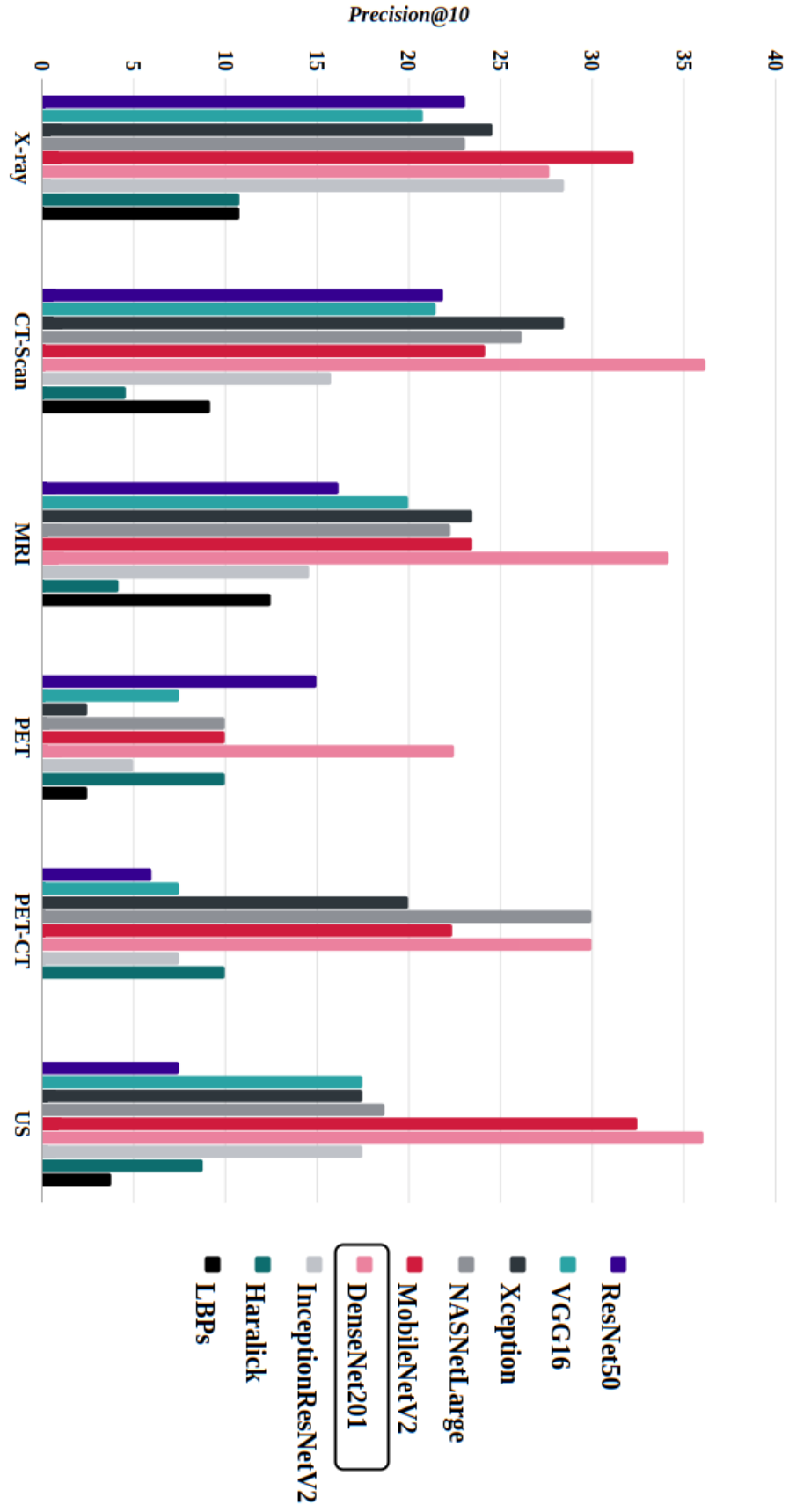


**Figure 6.4** Retrieval Performance: deep features (with random weights) vs texture features.

**Figure 6.5 Sample Query Image.**


**Figure 6.6 ResNet50: Top-5 Retrieval Results.**

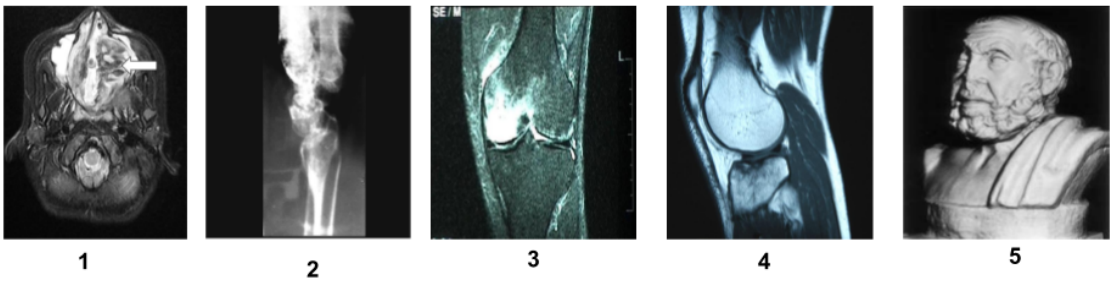


**Figure 6.7 DenseNet201: Top-5 Retrieval Results.**

that medical image content features might be helpful for modality-based retrieval. On the other hand, since query by example is the most basic form for expressing information needs by radiologists, we thought content-based retrieval offers a better starting point for a general-purpose medical image retrieval system as it alleviates the dependency on annotations that a text-based system would need which are nevertheless sometimes unreliable. Implementing content-based search in a general-purpose system also reduces engineering costs needed to adapt to different annotations procedures practiced by different healthcare institutions. Moreover, the constant change of imaging proto-

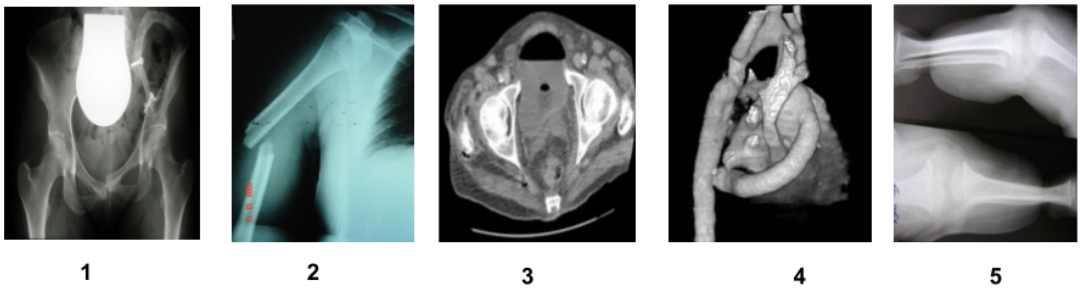


**Figure 6.8 Halarick: Top-5 Retrieval Results.**

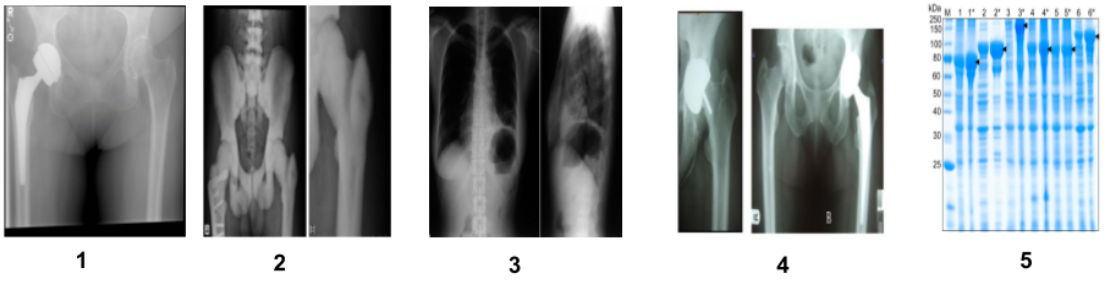


**Figure 6.9 LBPs: Top-5 Retrieval Results.**

cols makes it impossible to rely on DICOM tags to categorize medical images [GKK+02] automatically.

## 6.5 Related Work

In literature, most studies focusing on feature representations for medical images have analyzed specific features in a dataset with a single modality of medical images. With this setting, we can not know how valuable those features are when applied to other medical imaging modalities; therefore, using those features in a general-purpose medical image retrieval system is risky. One reason such studies are familiar is the difficulty of obtaining datasets containing various medical images due to confidentiality issues. The followings are some research works that studied texture features on specific modalities; X-rays [AKG+15; BTZ+17; RBH14; ZFL+12], MRI [DMM19], CT [XNH+18; PK11], PET-CT [SSG+19]. On the other hand, works that studied deep features include X-rays [SYS+17; AKG+16; AKG+15; LTK16; SPC+20], CT [HL15b; CLQ+19b; VWL+19] and MRI [SCN+16]. In contrast, our study examined each feature across different medical imaging modalities; hence we could evaluate the usefulness of these features for content-based retrieval for general-purpose medical image retrieval systems.

## 6.6 Discussion

In this chapter, we studied the retrieval of medical images based on their contents. We have looked into two kinds of feature representations, handcrafted features (mainly texture features) and deep features, and evaluated their performance in identifying similar images considering their semantics and modalities. We presented results showing the performance of texture features, namely Haralick's and LBPs textures and deep features by different CNNs architectures, namely, VGG16, InceptionResNetV2, ResNet50, DenseNet201, MobileNetV2, Xception, and NASNetLarge. Overall, the results show that deep features perform better than texture features, making them a better feature representation method for medical images. Furthermore, our study suggests that CNNs architectures with skip connection like ResNets, and DenseNets are better architectures for learning medical image features, making them suitable feature extractors for general-purpose medical image retrieval systems. On the other hand, our study shows the drawback of relying on contents alone, as sometimes the system retrieves insignificant random images, which does not add value in augmenting radiologists' diagnosis workflow. In the subsequent study, we address this limitation by supplementing image contents with patient demographics and disease predictions to inform retrieval.

# 7

# Retrieval Based on Contents, Patients' Demographics, and Disease Predictions

As we have seen in the previous study (Chapter 6), by depending solely on medical image contents, the system sometimes retrieves random images that are insignificant to the comparison analysis required by the radiologists. To improve that, we can take inspiration from how radiologist works in clinical settings. Usually, when diagnosing certain images, radiologists would consider other information available about the patients, including demographics, symptoms, laboratory data, and other necessary information to get the clinical context right. A medical image retrieval system also needs to leverage such information to identify the similarity between different cases of medical images effectively. This chapter presents a study in which we complement medical image contents with patient demographics to inform the retrieval. The radiologist can express information needs by submitting example images and patients' demographic information. We have also added a deep learning-based disease prediction model to assist the radiologists in image interpretation and help accurately identify and retrieve similar cases.

## 7.1 Introduction

Usually, when diagnosing certain medical images, radiologists would consider other information available about the patients, including demographics, symptoms, laboratory data, and other necessary information to understand the clinical context. This is important because such information can help tell the difference between cases. For example, in CXRs, visual patterns that indicate pneumonia would be precise in one patient with an elevated white blood cell count and fever. However, similar visual patterns for another

**Table 7.1 Public Datasets for CXRs**

| Dataset | Release year | # findings | # samples | Image-level labels | Local labels |
|---|---|---|---|---|---|
| JSRT [SKI+00] | 2000 | 1 | 247 | Available | Available |
| MC [JCA+14] | 2014 | 1 | 138 | Available | N/A |
| SH [JCA+14] | 2014 | 1 | 662 | Available | N/A |
| Indiana [DFKR+16] | 2016 | 10 | 8,121 | Available | N/A |
| ChestX-ray8 [WPL+17] | 2017 | 8 | 108,948 | Available | Available |
| ChestX-ray14 [WPL+17] | 2017 | 14 | 112,120 | Available | N/A |
| CheXpert [IRK+19] | 2019 | 14 | 224,316 | Available | N/A |
| Padchest [BPS+20] | 2019 | 193 | 160,868 | Available | N/A |
| MIMIC-CXR [JPB+19] | 2019 | 14 | 377,110 | Available | N/A |
| VinDr-CXR [NLL+22] | 2020 | 28 | 18,000 | Available | Available |

patient who doesn't have the same laboratory values and clinical characteristics might indicate other etiologies such as lung cancer, pulmonary edema, or atelectasis [HPS+20]. Demographic information, particularly Age, on the other hand, is critical information since "it is an essential factor that contributes to phenotypic changes in health and disease and, therefore, can affect the course and progression of a disease [MS22]". It is also vital in determining the correct course of treatment [GCR13]. A study by [BBS21] reveals significant age and gender-related (another demographic trait) differences in cardiac size parameters acquired from routine frontal CXRs. Furthermore, this study concluded that, if considered, those differences may result in appropriate and early intervention of cardiac pathology in some age groups. A medical image retrieval system must also consider additional information when comparing cases to identify their similarities or differences. This is important for almost all diagnostic radiology images but even more critical to medical imaging examinations that can screen more than one condition/disease when examining the same anatomical structure. With these kind of images, it means that even if the retrieval system gets the similarity based on the contents correct, there is still uncertainty; therefore, additional clinical information about the patient can make a difference. An excellent example of such examination is CXR. CXRs screens lung conditions, heart-related lung problems, the size and outline of the heart, the condition of blood vessels, calcium deposits, fractures, postoperative changes, e.t.c. and therefore it reports many diseases. An excellent example of this phenomenon is illustrated in Table 7.1 that shows some of the CXRs public datasets. Here, JSRT [SKI+00] dataset has only a single finding, while Padchest [BPS+20] contains 193 findings.

As we explained above, not only CXR examination can report multiple diseases but also other imaging modalities examining different parts of a human body. This means a general-purpose medical image retrieval system must handle this situation effectively

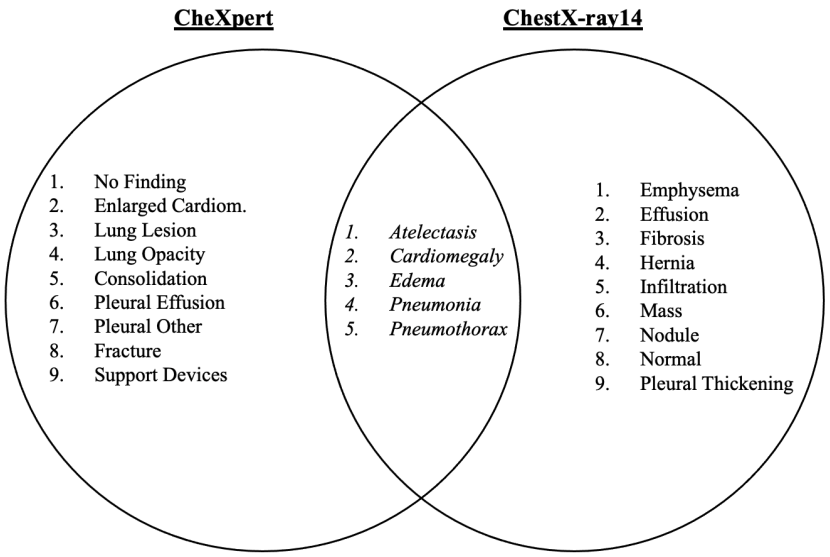


**Figure 7.1 Disease labels in CheXpert and ChestX-ray14 datasets.**

by allowing radiologists to express their information need by adding as much patient information as deemed necessary and using them to identify and retrieve similar images. However, with this additional information, the question remains; How can we leverage medical image contents and other patients' information to inform the retrieval process?

## 7.2 Methodology

### 7.2.1 Datasets

Due to the availability of the public dataset that contains medical images together with demographic information, we used two public CXRs datasets, CheXpert [IRK+19] and ChestX-ray14 [WPL+17] for this study. We mainly used a part of the former dataset to train the deep learning models for feature representation and disease predictions. While for the retrieval experiments, we used the former (20% held-out test set) and the latter. CheXpert is a publicly available dataset from Stanford University Hospital. It contains 224,316 CXRs of 65,240 patients automatically labeled with 14 diseases extracted from radiology reports. On the other hand, ChestX-ray14 comprises 112,120 images of 30,805 unique patients collected from 1992 to 2015 by the National Institutes of Health (NIH). These datasets have only five labels in common (see Figure 7.1).

### 7.2.2 The CheReS Approach

To complement image contents with patients' demographics, we propose a multi-faceted method we call *CheReS*. CheReS uses two deep learning models to analyze image contents. First, an Autoencoder model to create lower-dimensional feature representations of medical images. Second, a DenseNet model to predict diseases on the query images submitted by radiologists. On the other hand, CheReS uses a deterministic algorithm to integrate the content-based search results based on the autoencoder's feature representations, patient demographics, and disease predictions to retrieve the most accurate similar images eventually. Figure 7 shows CheReS' retrieval process from when a query is submitted to the final results. In the following sections, we give details on each of the CheReS' three components; *Feature Representation*, *Disease Predictions*, and the *Identification of Similar Cases*.

#### 7.2.2.1 Feature Representation

The CheReS' starting point for identifying similar cases relies on image contents, and therefore, we need a feature representation technique good enough for this purpose. We trained an autoencoder model, a deep neural network for dimensionality reduction and representation learning, and eventually use it as a feature extractor. Autoencoder mainly brings two advantages; First, it is an unsupervised learning approach, meaning it can learn distinct features for medical images without depending on their labels. Second, it can create lower-dimensional feature vectors, reducing computational costs during the similarity search process.

To make the representations produced by the autoencoder more efficient for the retrieval process, we took inspiration from Xie et al. [XGF16]. Xie et al. proposed an unsupervised deep embedding for the clustering analysis in which two phases are involved for the autoencoder to produce optimal feature representations that can cluster images based on the similarity of their contents. The first phase is parameter initialization, in here, we trained the model to compress each raw input image to a ten-dimensional vector using a reconstruction loss (see Figure 3.11). The second phase is parameter optimization, where the initial representation is refined for clustering using $KL$ divergence loss. During this step, we stacked the clustering layer on the encoder part of the model. Therefore, given input images, the model learns the similarity between them by clustering their latent representations into different clusters while fine-tuning the weights of the clustering layer and encoder mutually. This means the model learns to create representations of similar images to be closer to the embedding space.

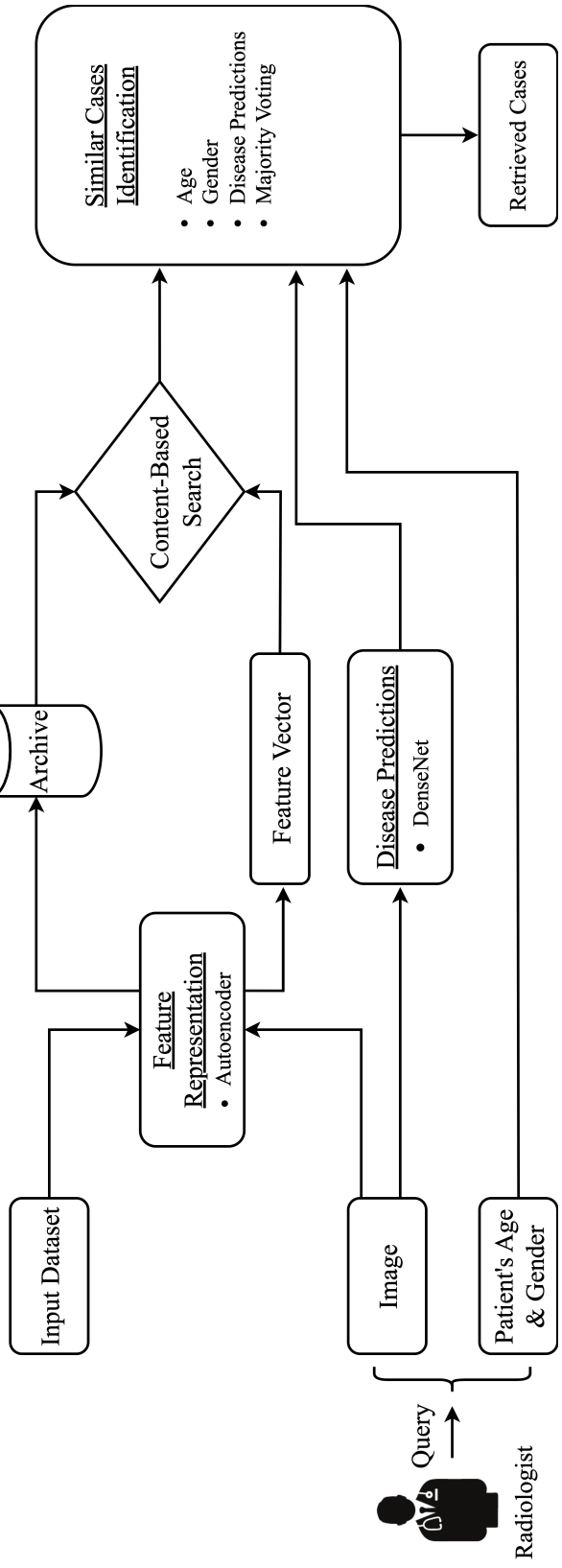


**Figure 7.2 CheReS' Retrieval Process.**

#### 7.2.2.2 Disease Predictions

For the disease predictions, we trained a DenseNet model to predict five diseases: Atelectasis, Cardiomegaly, Edema, Pneumonia, and Pneumothorax. We chose a DenseNet architecture in building our model owing to the lesson learned in Chapter 6 and further proved by [IRK+19] that this architecture is effective for learning abstract concepts in medical images. On the other hand, we trained the model to predict those five diseases because they are common chest diseases; therefore, their predictions can significantly inform the retrieval and help radiologists establish a diagnosis baseline. "Cardiomegaly, for example, and other heart diseases is the number one leading cause of diseases in the year 2019 in the United States, while Pneumonia is the $9^{th}$ most often diagnosed disease in the United States [CDC21]. Cardiomegaly can cause heart failure, and the likelihood of this correlates with age and is more common in males [TZZ+12]. Pneumonia, on the other hand, kills around 50,000 people in the United States alone, and more than 1 million adults are hospitalized yearly [RIZ+17]".

#### 7.2.2.3 Identification of Similar Cases

To identify similar cases, CheReS first searches for 50 images based on their contents and then refines the results to the ten most similar cases. We chose ten cases as the final results since it is less likely to bring information overload to radiologists, and it is feasible to compare them quickly. CheReS identifies these ten similar cases by including disease predictions on the image to be interpreted and the patient's demographic data (age and gender). We have also considered the frequency of diseases that appear the most in the first 50 images (majority voting) to refine the search results. The majority voting of diseases is based on the disease labels that are retrieved together with these 50 images from the archive. The intuition of adding majority voting is to accommodate new disease labels outside of five disease predictions. This makes CheReS robust to different datasets with different disease labels, another essential capability needed in a general-purpose system.

Algorithm 7.1 shows the step by step of the process. First, a ranked list $P_d$ based on probabilities of diseases predictions is created, followed by a ranked list $M_d$ of the five major diseases found in the top 50 images retrieved at first. "CheReS must then decide how to use predicted or major disease rankings at this stage. To do that, we calculated a Spearman Rank Correlation [TFE21] between the two ranked lists. Spearman Rank Correlation is calculated as follows;

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \tag{7.1}$$

Where $\rho$ is Spearman's Rank Correlation Coefficient, $d_i$ is the difference between the two ranks of each observation, and $n$ is the number of observations. The value of $\rho$ is inside the interval $[-1, 1]$ where 1 implies the agreements between two rankings is a perfect match. At the same time, $-1$ indicates the mismatch between the two ranked lists [TFE21]".

**Algorithm 7.1 Identification of Similar Cases**

**Input:** The image to be interpreted $I_q$
**Output:** Top 10 similar cases

Initialize query image $I_q$
Predict 5 diseases on query image: $P_d = DenseNet(I_q)$
Generate feature representations and retrieve $top50$ similar images based on $L_2$ distance
Get 5 diseases based on Majority Voting$M_d$ from $top50$
Calculate Spearman Correlation: $\rho = SpearmanR(P_d, M_d)$
**if** $\rho > 0$ **then**
    **if** $\rho == 1$ **then**
        Rank images based on $L_2$ distance and diseases in $P_d$
    **else**
        Rank images based on $L_2$ distance and positively predicted
        diseases in $P_d$ and $top2$ diseases in $M_d$
    **end if**
**else**
    Rank images based on $L_2$ distance and diseases on $M_d$
**end if**
Re-rank images based age and gender
Return $top10$ similar images.

"Suppose there is no match ($\rho < 0$), images with diseases in $M_d$. If $\rho = 1$, the algorithm ranks higher, images with lower $L_2$ distances, and have diseases similar to those in $P_d$. If there is a slight correlation, $1 > \rho > 0$, then the ranking is based on the top two diseases from $M_d$ and only positively predicted diseases (diseases with $> 50\%$ risk) from $P_d$. The final results are then re-ranked, giving higher preference to similarity based on age and gender groups in the context of diseases as defined by [GCR13]".

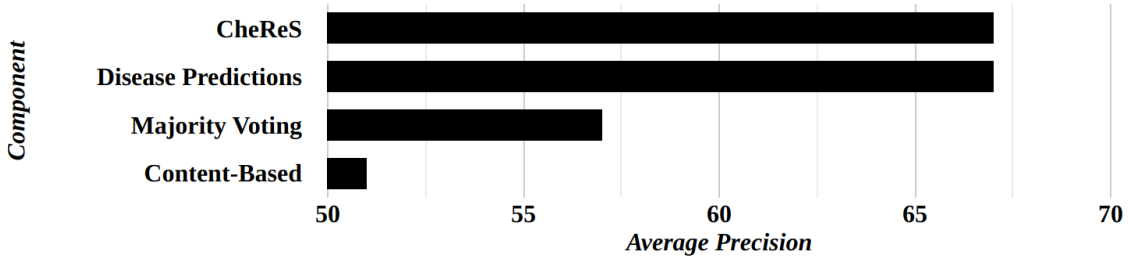


**Figure 7.3** **Retrieval perfomance for different components in the CheXpert dataset (CheReS refers to all components combined).**

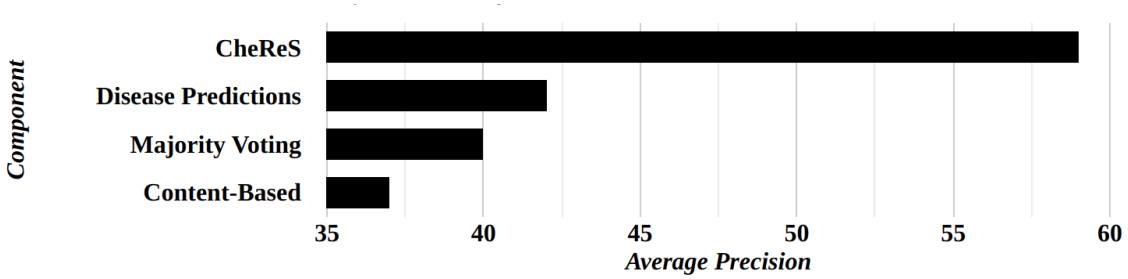


**Figure 7.4** **Retrieval performance for different components in the ChestX-ray14 dataset (CheReS refers to all components combined).**

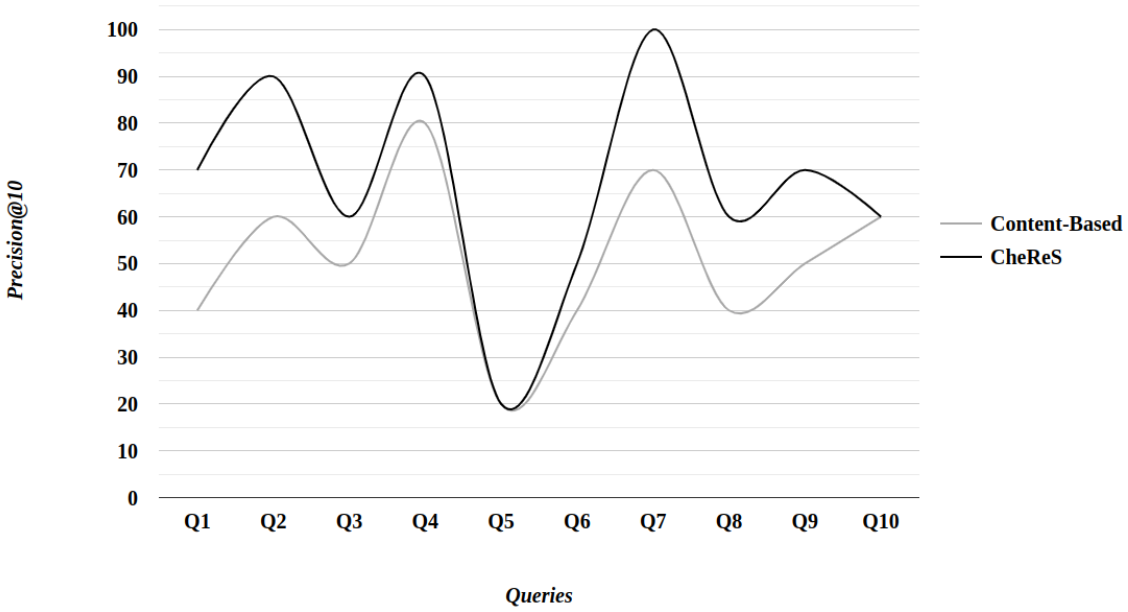


**Figure 7.5** **Retrieval performance for the random queries in the CheXpert dataset.**

## 7.3 Retrieval Performance

We evaluated the retrieval performance using the *precision*@10 and AP metrics. In each dataset, the retrieval was considered accurate if any of the images retrieved contained disease labels that overlapped with the query's ground truth diseases but also belonged to the same age and gender group of the patient. We conducted two kinds of evaluations.

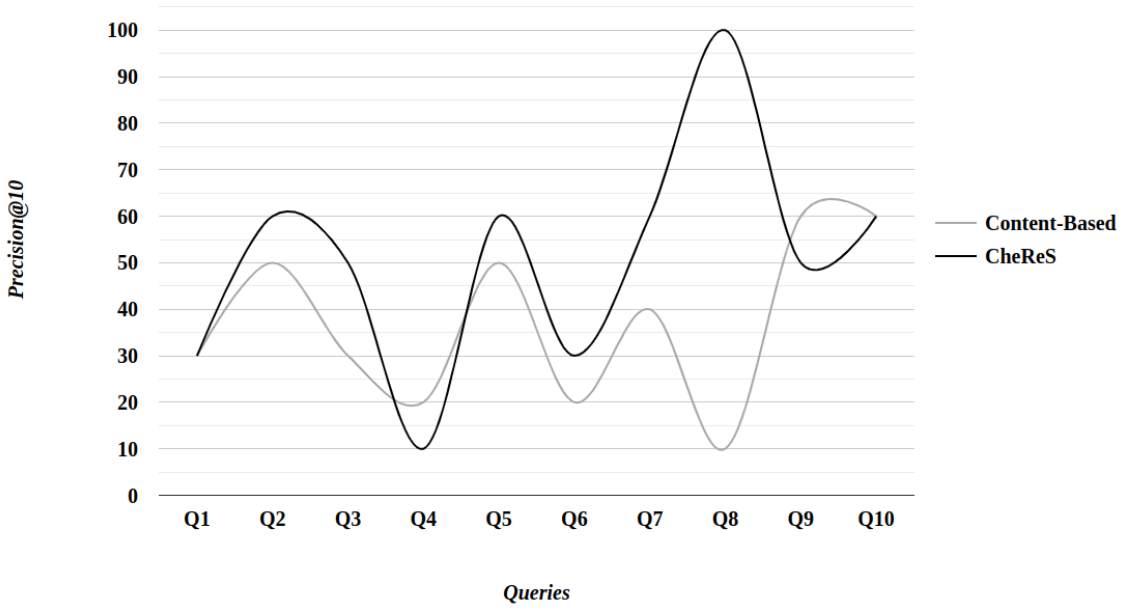


**Figure 7.6 Retrieval performance for the random queries in the ChestX-ray14 dataset.**

First, we wanted to check the contribution of each of the CheReS' components. This was done by adding one component after another to check the contribution to improving the retrieval results. As Figure 7.3 shows, we found that using disease predictions had a clear advantage (10% gain) compared to using the majority voting of diseases when the retrieval is conducted in the CheXpert dataset, which is the same dataset the disease prediction and feature representation models were trained on. However, in a completely different dataset (ChestX-ray14 dataset), using majority voting performed competitively to disease predictions. Here, disease predictions had only 2% gain(see Figure 7.4). This shows the contribution of considering the frequency of disease labels from the archive when applied to a different dataset and hence its importance as a component of the CheReS approach. Overall, the CheReS approach (with all its components) performs better than individual components in both datasets, showing its effectiveness compared to content-based approach.

We then randomly selected and also chose images with rare diseases to further evaluate CheReS. At this phase, we compared the overall CheReS method against the content-based approach. Figure 7.5 and 7.6 show the retrieval results. We observed that in the CheXpert dataset, CheReS outperformed the content-based approach for almost every query because of the significant help the disease predictions component provides. While on the ChestX-ray14 dataset, we found that the content-based approach still performed well in two queries ($Q4$ and $Q9$) with a rare disease called Hernia. Hernia was not among the five diseases the model trained to predict. By adding disease predictions and the majority voting of diseases from the archive, we added noise to the retrieval process,

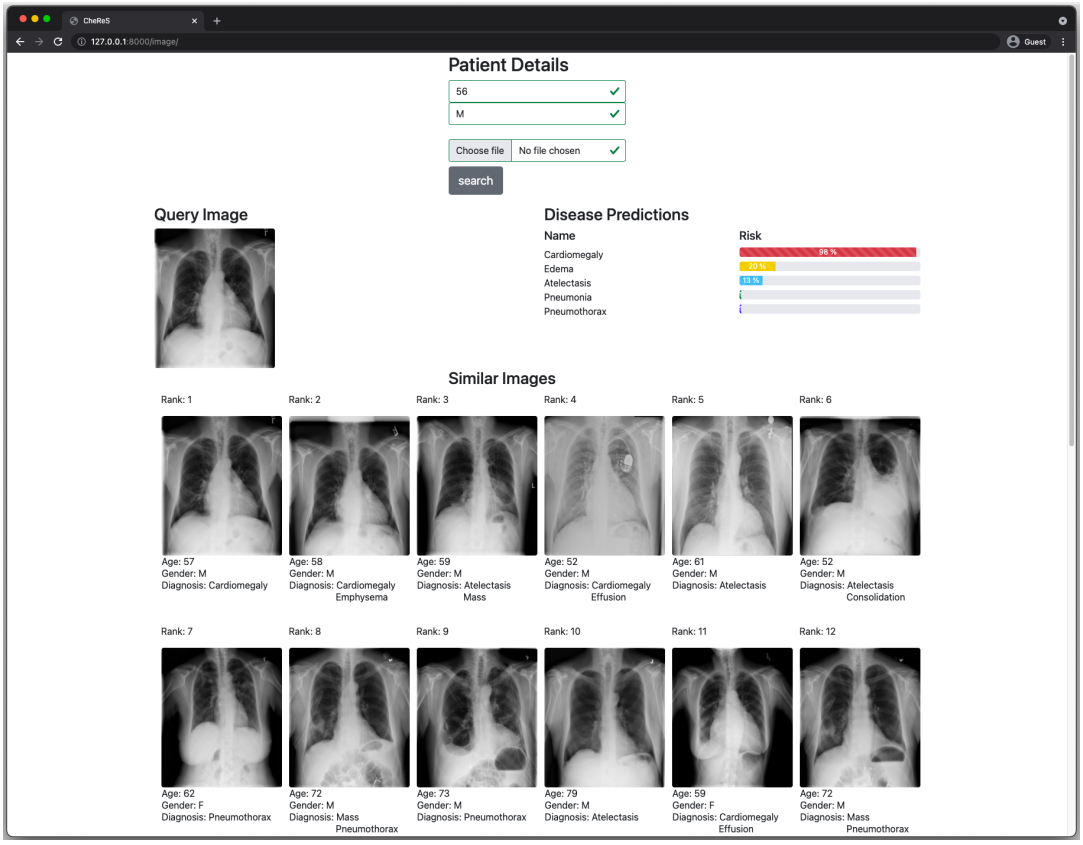


**Figure 7.7 CheReS' Interface.**

eventually affecting the retrieval quality. This shows that CheReS still has limitations in dealing with rare diseases and can use a room for improvement. To illustrate how radiologists can use the CheReS approach, Figure 7.7 shows the CheReS interface, in which disease predictions for the query image are displayed in the upper right while the bottom displays the retrieved images.

## 7.4 Summary

This chapter introduces CheReS, an approach to complement medical image contents with patient demographics when retrieving medical images. Furthermore, CheReS adds deep learning-based disease predictions to identify the similarity between different cases of medical images accurately. Our experimental results show that not only does CheReS significantly improves the retrieval results, but it is also robust enough for various datasets. However, one limitation of this study is that we have only used CXR datasets to prototype the CheReS approach. Nevertheless, the CheReS approach can be extended to any medical images; however, one has to choose which diseases should be included in the disease predictions part and how that information can be used to inform the identifica-

tion of similar cases process. The other limitation is that since CheReS allows radiologists to express their information need through sample image and patient demographics only, it is, therefore, most valuable when a radiologist openly searches for similar cases with the clinical context in mind. However, radiologists sometimes want to guide the retrieval process (especially when the visual clues in the image to be interpreted give them the indication of certain diseases) by targeting specific images that are significant in the comparison analysis needed to diagnose the image at hand. The following study presents an approach that guides the retrieval process by combining medical image contents with radiologists' text descriptions to target specific images to be retrieved.

# 8

# Retrieval Based on Contents and Radiologists' Text Descriptions

Instead of openly looking for similar cases, sometimes radiologists want to retrieve specific images. In other words, they want to guide the search process. This is because only some images are valuable to augment the comparative analysis that confirms or rules out initial hypothetical diagnoses (The differential diagnosis procedure). The previous approach (Section 7) is limited in accomplishing this need of radiologists because even though it improves the retrieval results compared to a purely content-based approach, it does not allow the radiologists to express their information needs by specifying what image should be retrieved that are significant for the comparison analysis needed for the differential diagnosis. In this chapter, we present a study on a guided search method that uses a deep metric learning technique and guarantees that radiologists retrieve significant images needed for the comparative analysis of the differential diagnosis.

## 8.1 Introduction

In healthcare, a differential diagnosis is a process of distinguishing a particular disease from others that present similar clinical features [TFE22b]. Usually, during this procedure, a clinician seeks to confirm or rule out a list of likely diagnoses before concluding with a definitive diagnosis. In interpreting medical images particularly, retrieving images with seemingly similar visual patterns could help radiologists in this procedure as it enables them to do a comparative analysis that can reveal other distinctive patterns that help confirm or rule out the initial hypotheses. To illustrate this more, consider Figure 8.1. Here, both images present the same visual patterns for consolidation. However, image number 6, at first glance, looks like consolidation, but if looked carefully, it is a nodular interstitial lung disease (Sarcoidosis) that is so widespread that it looks like con-

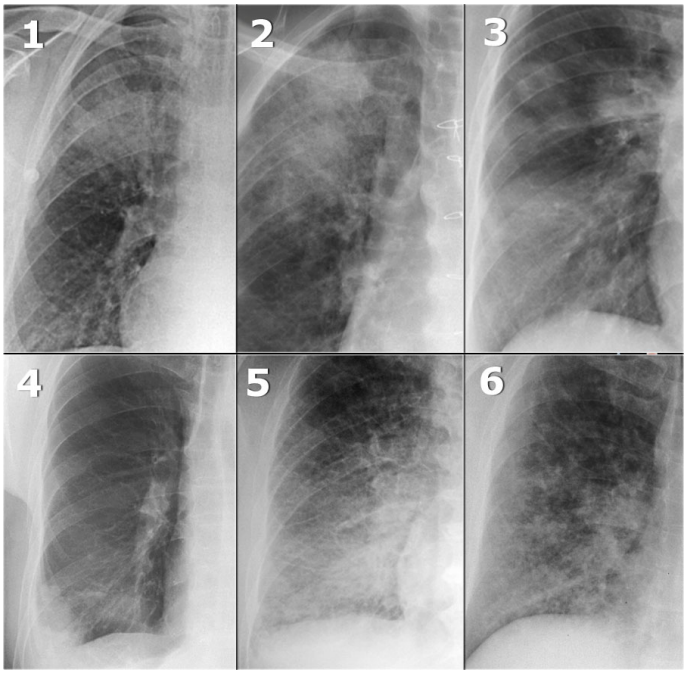


**Figure 8.1 CXR images with consolidation. The definitive diagnoses are: 1-Lobar Pneumonia, 2-Pulmonary Hemorrhage, 3-Organizing Pneumonia , 4-Infarction 5-Pulmonary Cardiogenic Edema and 6-Sarcoidosis [Smi22].**

solidation [Smi22]. If John, the radiologist, had to interpret this image and hypothesizes that it has either consolidation or sarcoidosis, it would be beneficial for him to compare it with images that have either sarcoidosis or consolidation to help clear the ambiguity and hence concludes that the image has sarcoidosis, a definitive diagnosis in this case.

To augment the differential diagnosis procedure, it would be beneficial if a medical image retrieval system utilized John's initial hypotheses by letting him submit the image to be interpreted and text descriptions to target specific images to be retrieved. However, one crucial challenge for this approach to work is fusing visual and text information to form the query. On the other hand, there is still a challenge in determining the similarity between this multimodal query and the required images.

## 8.2 Methodology

### 8.2.1 Problem Formulation

"We formulated the differential diagnosis task as a guided search problem to address the earlier-mentioned challenges. We mainly consider a scenario where a radiologist has an ambiguous image to interpret, and the visual clues suggest two diseases. To do a further comparative analysis, which could differentiate between these two initial hypotheses, a radiologist wants to search for confirmed images with similar visual patterns but has

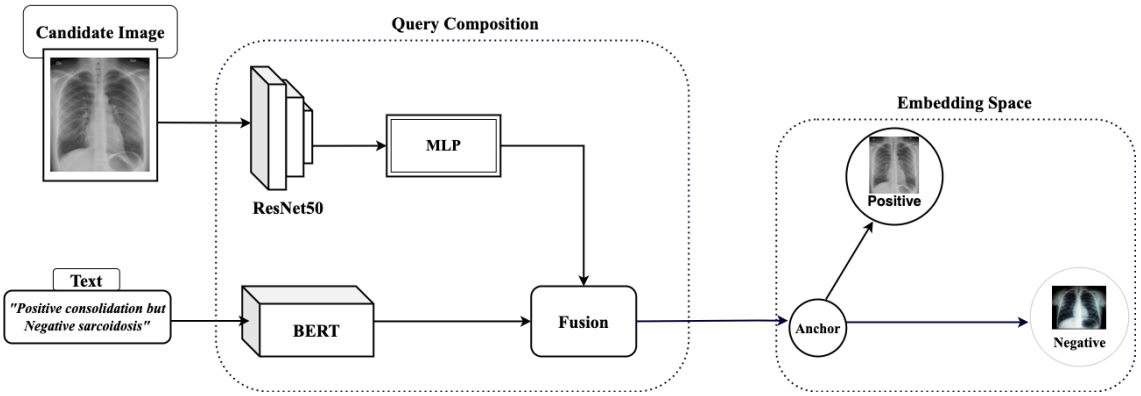


**Figure 8.2 A schematic overview of the proposed approach.**

been diagnosed with either. To achieve that, we want the radiologist to issue a query that comprises the image to be interpreted and the text like "Positive sarcoidosis but negative consolidation." In other words, this query instructs the system to "retrieve images similar to the query which have sarcoidosis but do not have consolidation" or vice versa if the text indicates otherwise. To combine image and text to form a query and also determine the similarity between this query and the target images to be retrieved, we propose a DML approach (refer Section 5.1.1.3) that learns to put this query closer to the target images and further from other images in the embedding space".

## 8.2.2 Proposed Approach

"Our proposed approach is depicted in Figure 8.2 and has two phases. In the first phase, we designed and trained a DML model to learn i.) how to compose query representations that leverage information from both image and text and ii.) the embedding function that places the query representations closer to the targeted images to be retrieved and far away from other images in the embedding space. In the second phase, we take images and texts from the test set, create their feature vectors using the trained model, and then perform a nearest neighbor search ($kNN$) to retrieve the required images for each query [MS21]".

### 8.2.2.1 Query Composition:

Owing to the lesson learned in Section 6.3, that ResNets are good CNN architectures to extract medical image features, "we used a pre-trained ResNet50 model as an encoder to create the representation of images. We also used the pre-trained Bidirectional Encoder Representations from Transformers (BERT) model to create text representation". BERT is a language model designed to pre-train deep bidirectional representations from an unlabeled text by conditioning both right and left contexts in all layers concur-

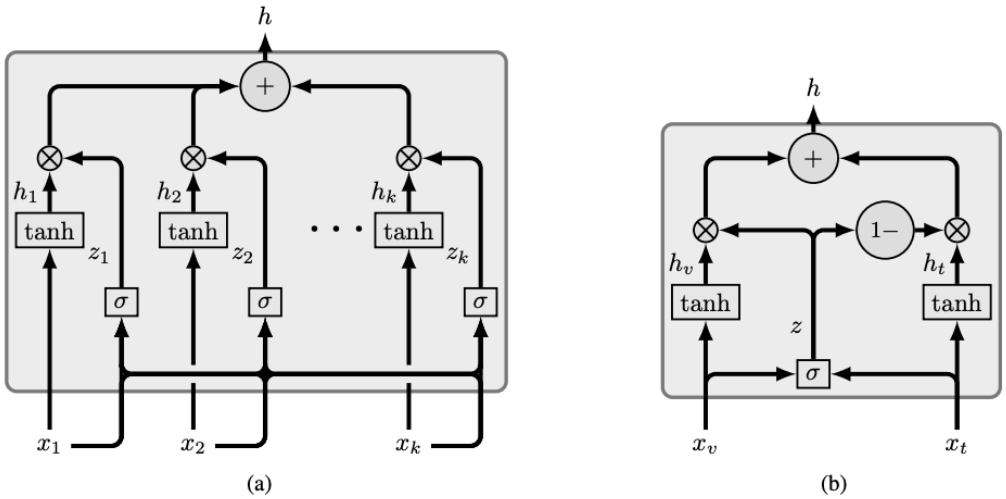


**Figure 8.3 Illustration of GMU. a) The model to handle more than two modalities. b) A simplified bimodal approach [OSM+17].**

rently [DCL+19]. By having a deep bidirectional architecture and attention mechanism, BERT learns the semantics of each word in a sentence but with all the nuances of context. BERT also considers the position of the words, which is essential for our method. For example, in the query text "positive consolidation but negative sarcoidosis," the position of the words positive and negative being prior to a particular disease is critical to the meaning of the whole query. "ResNet50 takes a raw input image of shape $(224, 224)$ and output a 2048-dimensional vector (after applying a global max-pooling in the last convolutional layer) while BERT encodes input texts to a 768-dimensional vector. To avoid these differences in dimensions between the two vectors influencing their positions during query composition, we first project the image through an MLP with 2048 and 768 neurons, respectively, to get image representation with a 768-dimensional vector similar to the text representation".

"To this point, both image and text representations are of equal dimensions and, therefore, can be passed through an image-text fusion model. We explored two methods of fusing images and texts to form a query representation. The first is simply concatenation in which the representation is formed through a concatenation operator: $M(x_v, x_t)$, where $M$ is an MLP. The other method regulates how information from either image or text contributes to the overall representation of the query. We designed a deep neural network model with Gated Multimodal Units (GMU). A GMU is an internal unit in a network architecture whose purpose is to find an intermediate representation based on an integration of data from different modalities [OSM+17]".

We used a simplified version of the model that handles two modalities (see Figure 8.3 b). Given feature vectors $x_v$ from the visual modality (image) and $x_t$ from text, each

vector feeds a neuron with a $tanh$ activation function with the goal to encode an internal representation $h_v$ and $h_t$ where $h_v = tanh(W_v \cdot x_v)$ and $h_t = tanh(W_t \cdot x_t)$. For each modality, there is a gate neuron ($\sigma$) which controls the contribution of features calculated from the inputs and thus makes the overall output of the unit to be $h = z \times h_v + (1-z) \times h_t$ where $z = \sigma(W_z \cdot [x_v, x_t])$ and $W_v$,$W_t$ and $W_z$ are parameters to be learned [OSM+17].

#### 8.2.2.2 Embedding Function

"As explained in Section 5.1.1.3, DML learns the embedding function that puts the representations of similar objects closer while dissimilar objects further apart in the metric space. For our case, we need the composed image-text query to be closer to targeted images with specific diseases while further away from other images. We can apply triplet loss [SKP15] for the learning process, and our composed query, targeted images, and other images will be anchor, positives, and negatives, respectively. A triplet loss ( 8.1) ensures that the similarity between the anchor and a positive pair $(a, p)$ is greater compared to the anchor and a negative pair $(a, n)$ by increasing and decreasing the distance between the pair of objects respectively. The margin value $\alpha$ acts as a threshold".

$$T_{loss} = \max(0, ||a - p|| - ||a - n|| + \alpha) \tag{8.1}$$

"During triplet selection, a common practice involves choosing the positives from the same labels with the anchors while the negatives are selected from other labels. This approach, however, is optimal where the labels are accurate, for example, in face verification or other related applications [SKP15], even if a positive pair of a person with different illumination or pose variances is chosen, it is still the same person. In medical images, depending entirely on labels is risky for the following reasons: i.) the available labels might not always be accurate because not all diseases are reported in an image due to omissions or when deemed not crucial by the radiologist [AM18]; and ii.) even if all labels would be present, their validity is still questionable because different diseases can adopt similar visual patterns and are still interpreted differently by different radiologists [CWS+19]. This means the validity of labels highly depends on how good a concerned radiologist is. With all these uncertainties, we need a loss function that can consider this. Thus, we decided also to explore a $MS_{loss}$";

$$MS_{loss} = \frac{1}{m}\sum_{i=1}^{m}\left\{\frac{1}{\alpha}\log\left[1 + \sum_{k\in P_i} e^{-\alpha(S_{ik}-\lambda)}\right] + \frac{1}{\beta}\log\left[1 + \sum_{k\in N_i} e^{-\beta(S_{ik}-\lambda)}\right]\right\} \tag{8.2}$$

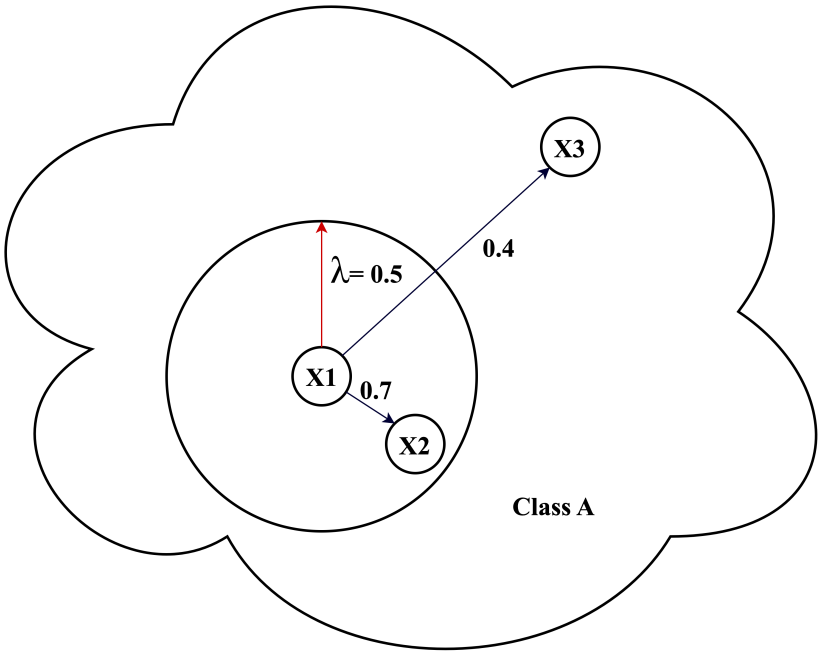


**Figure 8.4 Illustration of intra-similarity between anchor and positive samples in $MS_{loss}$ (x1=anchor, x2, x3=positives, $\lambda$=margin) [G20].**

where $s_i$ is the cosine similarity between pairs, $\lambda$ is a similarity margin and $\alpha, \beta$ are hyperparameters [WHH+19]. $MS_{loss}$ comprises positive and negative parts that deal exclusively with positive and negative pairs. For the positive part, given an anchor with $k$ positives, positive pairs whose similarities are less than $\lambda$ are heavily penalized (distant positives), making their loss higher than closer positives (those with similarities higher than $\lambda$). As Figure 8.4 illustrates, there are two pairs [x1:x2] and [x1:x3], positive part of the loss for [x1:x2] would be minimal since $e^{-\alpha(S_{i_k}-\lambda)} = e^{-\alpha(0.7-0.5)} = e^{-0.2\alpha}$; since $\alpha$ is hyper-parameter and usually higher than zero, value of this term will be lower when compared with positive part of the loss for [x1:x3]. For this pair, loss will be $e^{-\alpha(0.4-0.5)} = e^{0.1\alpha}$. A clear distinction appears among the loss back propagated for [x1:x2] and [x1:x3].

The negative part ensures that negatives have as minimal as possible similarity with the anchor by heavily penalizing the negative pairs closer to the anchor than those further away as Figure 8.5 shows. In here, the loss for [x1:x2] is $e^{\beta(S_{i_k}-\lambda)} = e^{\beta(0.3-0.5)} = e^{-0.2\beta}$, whereas loss for [x1:x3] is $e^{\beta(0.1-0.5)} = e^{-0.4\beta}$, since $e^{-0.2\beta} > e^{-0.4\beta}$ for $\beta > 0$ therefore the loss for negative pairs with higher similarity will be greater than negative pairs with low similarity [G20]. By considering these intra-similarities between anchors with positive and negative samples, $MS_{loss}$ reduces the uncertainties of depending entirely on labels.

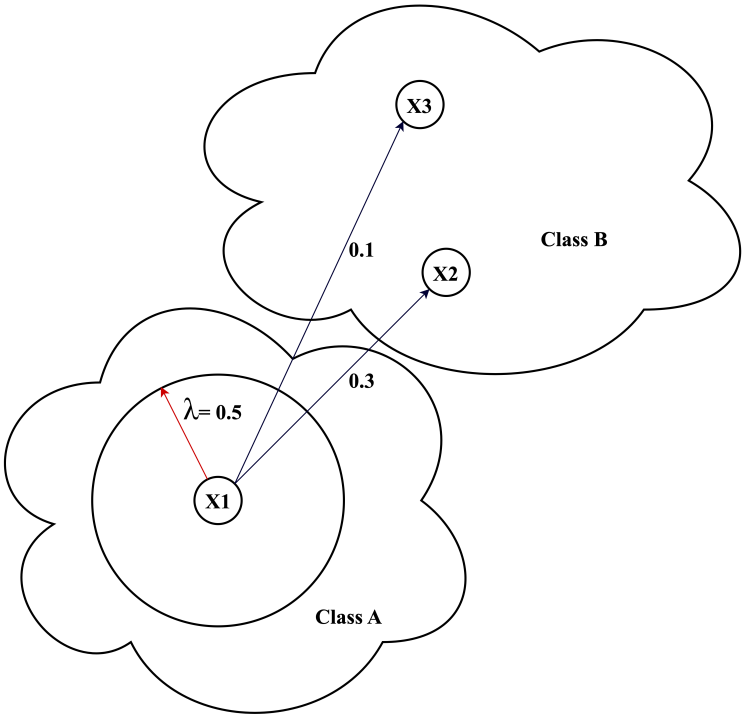


**Figure 8.5 Illustration of intra-similarity between anchor and negative samples in $MS_{loss}$ (x1=anchor, x2, x3=negatives, $\lambda$=margin) [G20].**

#### 8.2.2.3 Dataset

We used the CheXpert dataset to evaluate our proposed approach. As explained in Section 7.2.1, CheXpert contains 224,316 CXR images of 65,240 patients automatically labeled with 14 diagnoses extracted from radiology reports. Typically, such radiology reports are semi-structured documents in which radiologist record their interpretation in titled sections. The *findings* section contains a natural language description of the vital aspect of the image. In the *impression* section, a narrative summary of the most immediately relevant findings is given. In CheXpert, the 14 diagnosis labels were extracted from the impression section [IRK+19] using Natural Language Processing (NLP) techniques. We selected anchors, the images diagnosed with two diseases that appeared together not $< 5K$ times. With this setting, we ended up with $25K$ images with the following pair of diagnoses: i.) Edema and Atelectasis, ii.) Atelectasis and Lung Opacity, iii.) Lung Opacity and Consolidation, iv.) Pleural Effusion and Lung Opacity, and v.) Edema and Lung Opacity. For each anchor, we selected corresponding images with either of the two diagnoses as positives while other images as negatives.

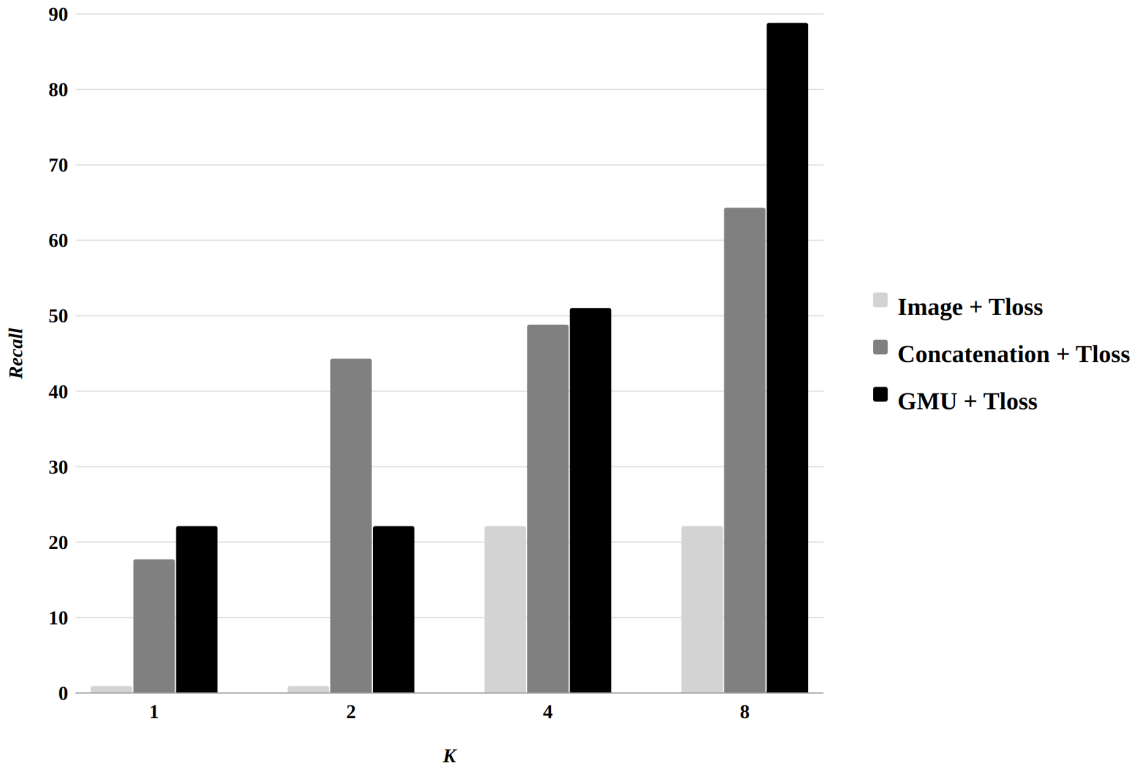


**Figure 8.6 Retrieval performance of different pipelines trained with triplet loss.**

## 8.3 Retrieval Performance

We trained models with concatenation and GMU image-text composition pipelines and a pipeline with only images as queries . Said differently, former pipelines combine image and text to form a query, while the latter relies entirely on image contents. For each pipeline, we also trained with triplet and multi-similarity losses. The models created were then used in the retrieval experiments to create the representations of queries and images to be searched through, both of which were from a held-out test set. The retrieval process was done through the nearest neighbor search using the $L_2$ distance metric and Recall@K as the performance measure. The retrieval results can be seen in Figure 8.6 and 8.7. As shown in both figures, no matter the loss function used, a DML approach that combines image and text to form queries outperforms image only approach (content-based). This shows the efficiency of our proposed approach that leverages information from both image and text when creating a query rather than relying on image contents only.

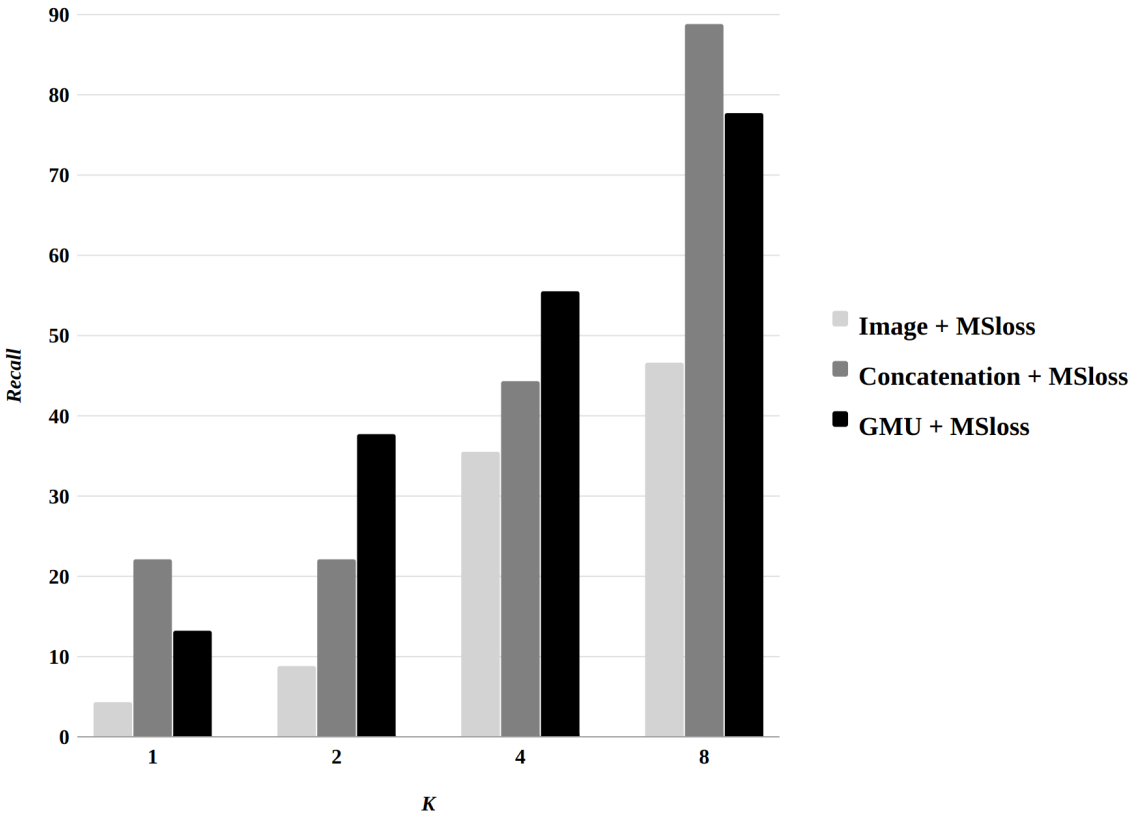


**Figure 8.7 Retrieval performance of different pipelines trained with multi-similarity loss.**

## 8.4 Discussion

Our study limited the radiologist to express their information need through images, and the text contains disease labels available in the CheXpert dataset. In other words, our proposed approach is as good as the accuracy of these disease labels. We understand from the previous study that adding more information like patient demographics, clinical history, symptoms, or lab tests could improve differential diagnosis outcomes. However, that would raise a new question on how to fuse all of this information in a DML process. Nevertheless, our approach has proven superior to image content alone. This is essential as a proof-of-concept in our approach to augment the differential diagnosis of medical images. On the other hand, we hope these promising results can inspire the need to explore different methods to augment differential diagnoses of medical images and therefore support radiologists in their daily workflow rather than trying to replace them. Only by doing so can we improve their diagnostic accuracy and increase their throughput [RHP+22].

## 8.5 Related Work

Most of the previous approaches in the literature do not directly address the differential diagnosis problem. One reason is that they assume the results of any retrieval system can allow a radiologist to do the comparative analysis needed for differential diagnosis. However, this is not the case; as we have seen in the retrieval evaluation of Chapter 6, a content-based alone approach sometimes retrieves insignificant random images. On the other hand, few available approaches in the literature include [LRL+19], and [LJE+19]. "[LRL+19] explored differential diagnosis for pancreatic cysts in CT images using a DenseNet model that learns high-level features from the entire abnormal pancreas to create the mapping between the medical imaging appearance and the various pancreatic cysts pathologies, and generate saliency maps to visualize essential regions. With this approach, a radiologist can only get information on potential diagnoses with respect to regions within the same image. However, he/she cannot compare this image with others, nor can he/she specify what diseases the other images should have when retrieving them".

On the other hand, Yuan et al. [LJE+19] took a bottom-up approach to the differential diagnosis of skin diseases. To augment the ability of general practitioners who did not have additional specialized training to diagnose skin conditions accurately, she created a deep learning-based system that distinguishes between 26 of the most common skin conditions and suggests potential diseases to be considered when diagnosing a patient. Unlike those studies, our work explored CXR images. CXR is the most common medical imaging examination globally [IRK+19], giving us a perfect sample to prototype our method for a general-purpose medical image retrieval system.

## 8.6 Summary

In this chapter, we proposed a new approach to augment the differential diagnosis of medical images based on DML. We have formulated the differential diagnosis task as a guided search problem in which a radiologist can issue a query combining the to-be-interpreted medical image and the text targeting to retrieve images with specific diseases that would significantly help the comparative analysis needed for the differential diagnosis procedure. We trained a DML model to compliment image contents with radiologists' text descriptions when forming a query representation. This model also learns the embedding function that puts the query closer to images with required diseases while further apart from other images in the embedding space. Compared to the content-based only approach that sometimes retrieves insignificant random images, this

approach guarantees that the retrieved images have similar visual patterns. However, they are with specific diseases expressed by the radiologist's text and, therefore, can help them make the comparative analysis needed for the differential diagnosis procedure. The limitation of this study, however, is that we should have included information about a patient, like symptoms, lab tests, etc.; however, that would require a new approach to fuse all this information, which is still an open question to be explored. Nevertheless, this is an exciting direction of research as the outcomes of it can make general-purpose medical image retrieval systems handle radiologist information need differently than traditional approaches and significantly augment their daily workflow.

Part IV

# Conclusion and Future Work

# 9

# Conclusion

The thesis contributes to the overall goal of building a general-purpose medical image retrieval system that can handle various information needs of radiologists through different query forms and thus augments them in their daily workflow, especially during the interpretation of medical images. The underlying motivation of this work is that building such a system will allow radiologists to search for similar cases to make comparative analyses that inform their diagnostic decisions. Consequently, that would help reduce diagnostic errors and significantly improve their throughput since they can interpret more images at a time, hence improving the lives of many patients [RHP+22].

Developing such a system, however, intensely requires collaborative efforts across multiple stakeholders, including computer scientists and clinicians. It requires building complex infrastructure and generating new security and privacy regulations across hospitals, academic research institutes, and multi-national consortia [ZGD+20]. Currently, the challenge of obtaining the required data from healthcare institutions hinders the progress of such endeavors. However, as more and more data becomes available, researchers will continue exploring innovative ways to contribute to creating general-purpose medical image retrieval systems. In this thesis, the availability of datasets partly limited the type of studies we could conduct. Nevertheless, we proposed fundamental retrieval approaches that we could address with the available datasets; however, their solutions are applicable across different medical images, making them suitable for integration into a general-purpose medical image retrieval system.

First, we studied the retrieval of medical images based solely on their contents in which a radiologist expresses information need by submitting a sample image (query by example). Then the system computes the similarities of medical images based on their contents. To find an optimal feature representation method that distinguishes medical images considering their semantics and modalities, we thoroughly analyzed different feature representation techniques based on handcrafted methods (mainly texture fea-

tures) and deep learning (deep features) to represent the contents of medical images. Based on the experiment results, we propose deep features as the optimal feature representation approach. On the other hand, we propose that deep CNNs with skip connection, ResNets and DenseNets are better architectures for learning and extracting medical image features. This result is significant because it guides what an effective feature extractor to deploy in a general-purpose retrieval system is.

Second, we studied the retrieval of medical images based on content and patients' demographics and proposed a multi-faceted method to accurately retrieve similar cases of medical images while considering the clinical context. Our approach allows radiologists to express information needs by submitting a query comprising a sample image and demographic information about the patient. This approach leverages both this information and deep learning-based disease predictions to understand the clinical context, making it able to identify similar cases accurately. This approach significantly improves the retrieval results and is robust enough to be used in different datasets, making it a perfect tool to be integrated into a general-purpose retrieval system.

Lastly, we studied the retrieval of medical images based on contents and text descriptions provided by radiologists. Here, we propose a method that allows radiologists to target specific images to retrieve by submitting a query consisting of a sample image and text descriptions expressing their information needs. This kind of search aligns well with the fact that radiologists are trained professionals, so when they look at the medical images, they form initial hypotheses before finding the definitive diagnoses. Said differently, this method allows radiologists to guide the search process. Guided search is one of the essential features needed in a general-purpose medical image retrieval system. Unlike traditional content-based approaches that rely on image features only, which sometimes retrieve insignificant random images, this guided search method combines an image with a radiologist's text description to guarantee that the retrieved images are suitable for the comparative analysis needed to confirm or rule out initial hypotheses (differential diagnosis). Having this capability in a general-purpose medical image retrieval system would significantly help radiologists in their daily workflow since differential diagnosis is one of the most complex tasks for radiologists, yet the most important one.

The studies in this thesis contribute a small step toward developing a general-purpose medical image retrieval system. However, much work is still needed to eventually make a perfect system that augments radiologists in their daily workflow. We envision a future where all information regarding a patient, including images, laboratory data, symptoms, and others, are integrated. This will allow the radiologist a wide range of possibilities for expressing their information needs when looking for similar cases to supplement

their diagnostic processes.

# 10

# Future Work

As explained earlier, this thesis contributes a small step toward developing a general-purpose medical image retrieval system. However, much work is needed to develop a viable system that augments radiologists in their daily workflow. In the following sections, we will outline our suggestions for the future work in this area.

## 10.1 Pixel-level Image Analysis

Due to the lack of medical image datasets that are well labeled, the studies in this thesis are based on image-level information only, like disease labels. However, the pixel-level analysis would give radiologists a detailed image analysis and help them interpret and compare medical images efficiently. For example, one can train a self-supervised model that learns the fine-grained image similarity and infers similarity between images from one pixel to pixel, thus understanding the regions of interest better. Deep neural networks based on the Transformer architecture [VSP+17], provide an attention mechanism that might be suitable for this task.

## 10.2 Data Integration

According to a study by Kroth et.al [KMDV+19], poor access to information the clinicians need during their practices contributes to stress and burnout. On the other hand, other studies have shown that most radiologists prefer access to data, like clinical and laboratory data, during medical image interpretation, as the lack of such access considerably affects their interpretation accuracy [HPS+20]. One reason, however, for the existing poor access is that clinical data collection, storage, and integration is still broken in healthcare. Typically, many healthcare organizations implement multiple Electronics Health Records (EHR) and other data collection systems maintained by different teams,

departments, etc., making it difficult for a particular patient's data to be integrated into a single point. Even more challenging is that these departments/organizations sometimes have different annotation practices, making data interoperability difficult. For accurate interpretation and similarity search of medical images, there is a need to integrate an image with all supporting information, from patient records to additional descriptors (such as vital signs, blood tests, medications, genomics, and non-imaging data such as ECG) [ZGD+20; HPS+20] to infer the clinical context. This means there is a need for optimal methods that ensure effective data integration for similarity search purposes while considering the clinical contexts.

## 10.3 Guided Search

Guided search can significantly reduce radiologists' time to find the necessary information, fasten their diagnosis process, and save lives, especially in critical situations. On the other hand, a guided search can help radiologists search for images with more detailed information that would vastly improve the quality of their comparative analysis. We envision the future where a medical image retrieval system can allow Johh, the radiologist search with a query that contains a to-be-interpreted chest CT image (as an example) that has ground-glass opacities and a text that guides the search with more fine-grained information, e.g., *Find images like this of a 60-year-old female with high-resolution, extensive patchy exudates of both lungs, faint ground-glass opacities on the edge, and interlobular septal thickening [DZY+20]*. Data integration is one step to enable the system to process such a query. The other step is designing effective machine learning methods to learn both image and text representations, considering the semantic contexts within and IR methods to index these representations to ensure accurate and faster retrieval.

# Bibliography


[Agr21] Aakash Agrawal. The why and the how of deep metric learning. 2021. URL: https://towardsdatascience.com/the-why-and-the-how-of-deep-metric-learning-e70e16e199c0. Visited on 09/05/2022.

[Aiz03] Akiko Aizawa. An information-theoretic perspective of tf–idf measures. *Information Processing & Management*, 39(1):45–65, 2003.

[AA18] Afshine Amidi and Shervine Amidi. Convolutional neural networks cheatsheet. 2018. URL: https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-convolutional-neural-networks. Visited on 25/07/2022.

[AKG+15] Yaron Anavi, Ilya Kogan, Elad Gelbart, Ofer Geva, and Hayit Greenspan. A comparative study for chest radiograph image retrieval using binary texture and deep learning classification. In *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2015, Milan, Italy, August 25-29, 2015*, pages 2940–2943. IEEE, 2015.

[AKG+16] Yaron Anavi, Ilya Kogan, Elad Gelbart, Ofer Geva, and Hayit Greenspan. Visualizing and enhancing a deep learning framework using patients age and gender for chest x-ray image retrieval. In Georgia D. Tourassi and Samuel G. Armato III, editors, *Medical Imaging 2016: Computer-Aided Diagnosis, San Diego, California, United States, 27 February - 3 March 2016*, volume 9785 of *SPIE Proceedings*, page 978510. SPIE, 2016.

[AM18] Mauro Annarumma and Giovanni Montana. Deep Metric Learning for Multi-Labelled Radiographs. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing (SAC 2018)*, pages 34–37, Pau, France. ACM, April 2018.

[AAK+19] Swarnambiga Ayyachamy, Varghese Alex, Mahendra Khened, and Ganapathy Krishnamurthi. Medical image retrieval using resnet-18. In *Medical Imaging 2019: Imaging Informatics for Healthcare, Research, and Applications*, volume 10954, page 1095410. International Society for Optics and Photonics, 2019.

[BKS+17] Morteza Babaie, Shivam Kalra, Aditya Sriram, Christopher Mitcheltree, Shujin Zhu, Amin Khatami, Shahryar Rahnamayan, and Hamid R. Tizhoosh. Classification and retrieval of digital pathology scans: A new dataset. *CoRR*, abs/1705.07522, 2017.

[BTZ+17] Morteza Babaie, Hamid R. Tizhoosh, Shujin Zhu, and M. E. Shiri. Retrieving Similar X-ray Images from Big Image Data using Radon Barcodes with Single Projections. In *Proc. ICPRAM*, pages 557–566, Porto, Portugal. SciTePress, February 2017.

[BDW+15] Yaniv Bar, Idit Diamant, Lior Wolf, and Hayit Greenspan. Deep learning with non-medical training used for chest pathology identification. In *Medical Imaging 2015: Computer-Aided Diagnosis*, volume 9414, page 94140V. International Society for Optics and Photonics, 2015.

[BGA+10] Ajay Basavanhally, Shridar Ganesan, Shannon Agner, James Monaco, Michael D. Feldman, John Tomaszewski, Gyan Bhanot, and Anant Madabhushi. Computerized image-based detection and grading of lymphocytic infiltration in HER2+ breast cancer histopathology. *IEEE Trans. Biomed. Eng.*, 57(3):642–653, 2010.

[BDE+17] Maike Baues, Anshuman Dasgupta, Josef Ehling, Jai Prakash, Peter Boor, Frank Tacke, Fabian Kiessling, and Twan Lammers. Fibrosis imaging: current concepts and future directions. *Advanced drug delivery reviews*, 121:9–26, 2017.

[BCV13] Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, 2013.

[Ber96] Leonard Berlin. Malpractice issues in radiology. perceptual errors. *AJR. American journal of roentgenology*, 167(3):587–590, 1996.

[BBH+12] J Biederer, M Beer, W Hirsch, J Wild, M Fabel, M Puderbach, and EJR Van Beek. Mri of the lung (2/3). why… when… how? *Insights into imaging*, 3(4):355–371, 2012.

[BBS21] Edmund Kwakye Brakohiapa, Benard Ohene Botwe, and Benjamin Dabo Sarkodie. Gender and age differences in cardiac size parameters of ghanaian adults: can one parameter fit all? part two. *Ethiopian Journal of Health Sciences*, 31(3), 2021.

[BCG18] Lindsay P Busby, Jesse L Courtier, and Christine M Glastonbury. Bias in radiology: the how and why of misses and misinterpretations. *Radiographics*, 38(1):236–247, 2018.

[BPS+20] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-Vayá. Padchest: a large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020.

[CWS+19] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. "Hello AI" Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proc. ACM Hum. Comput. Interact.*, 3(CSCW):104: 1–104: 24, 2019.

[CLW+10] Weidong Cai, Sidong Liu, Lingfeng Wen, Stefan Eberl, Michael J. Fulham, and David Dagan Feng. 3d neurological image retrieval with localized pathology-centric cmrglc patterns. In *Proceedings of the International Conference on Image Processing, ICIP 2010, September 26-29, Hong Kong, China*, pages 3201–3204. IEEE, 2010.

[CLQ+19a] Yiheng Cai, Yuanyuan Li, Changyan Qiu, Jie Ma, and Xurong Gao. Medical image retrieval based on convolutional neural network and supervised hashing. *IEEE Access*, 7:51877–51885, 2019.

[CLQ+19b] Yiheng Cai, Yuanyuan Li, Changyan Qiu, Jie Ma, and Xurong Gao. Medical Image Retrieval Based on Convolutional Neural Network and Supervised Hashing. *IEEE Access*, 7:51877–51885, 2019.

[CLM+11] Yu Cao, Y Li, H Müller, CE Kahn Jr, and E Munson. Multi–modal medical image retrieval. In *SPIE Medical Imaging*. Citeseer, 2011.

[CBL+04] Gabriella Castellano, Leonardo Bonilha, LM Li, and Fernando Cendes. Texture analysis of medical images. *Clinical radiology*, 59(12):1061–1069, 2004.

[Cat22] Prof. Dr. Philippe Cattin. Principles of medical imaging. Lecture slides, 2022. University of Basel.

[CDC21] CDC. Leading causes of death. 2021. URL: https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm. Visited on 26/10/2021.

[CDC22] CDC. Radiation in healthcare:x-rays. 2022. URL: https://www.cdc.gov/nceh/radiation/x-rays.html. Visited on 1/10/2022.

[CL17] Moitreya Chatterjee and Yunan Luo. Similarity learning with (or without) convolutional neural network. Lecture slides, 2017. Cutting-Edge Trends in Deep Learning and Recognition, University of Illinois.

[Cho17] François Chollet. Xception: deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1800–1807. IEEE Computer Society, 2017.

[CW17] Yu-An Chung and Wei-Hung Weng. Learning deep representations of medical images using siamese cnns with application to content-based image retrieval. *arXiv preprint arXiv:1711.08490*, 2017.

[Cli22] Florida Medical Clinic. What is diagnostic radiology and what is it used for. 2022. URL: https://www.floridamedicalclinic.com/blog/what-is-diagnostic-radiology/#:~:text=Diagnostic\%20radiology\%20refers\%20to\%20the,Ultrasound. Visited on 04/07/2022.

[Cra09] Nick Craswell. *Mean reciprocal rank*. In *Encyclopedia of Database Systems*. LING LIU and M. TAMER ÖZSU, editors. Springer US, Boston, MA, 2009, pages 1703–1703. ISBN: 978-0-387-39940-9.

[Cre00] Fabio Crestani. Neural relevance feedback for information retrieval. In *Uncertainty in intelligent and information systems*, pages 197–208. World Scientific, 2000.

[DZY+20] Wei-cai Dai, Han-wen Zhang, Juan Yu, Hua-jian Xu, Huan Chen, Si-ping Luo, Hong Zhang, Li-hong Liang, Xiao-liu Wu, Yi Lei, et al. Ct imaging and differential diagnosis of covid-19. *Canadian Association of Radiologists Journal*, 71(2):195–200, 2020.

[DSH+19] Mohammad I Daoud, Amro Saleh, Ismail Hababeh, and Rami Alazrai. Content-based image retrieval for breast ultrasound images using convolutional autoencoders: a feasibility study. In *2019 3rd International Conference on Bio-engineering for Smart Technologies (BioSMART)*, pages 1–4. IEEE, 2019.

[HKD+13] Alba Garcia Seco de Herrera, Jayashree Kalpathy-Cramer, Dina Demner-Fushman, Sameer K. Antani, and Henning Müller. Overview of the ImageCLEF 2013 Medical Tasks. In *Working Notes for CLEF 2013 Conference*, volume 1179, Valencia, Spain. CEUR-WS.org, 2013.

[VWL+19] Bob D. de Vos, Jelmer M. Wolterink, Tim Leiner, Pim A. de Jong, Nikolas Leßmann, and Ivana Isgum. Direct automatic coronary calcium scoring in cardiac and chest CT. *IEEE Trans. Medical Imaging*, 38(9):2127–2138, 2019.

[DAED+15] Maria De-Arteaga, Ivan Eggel, Bao Do, Daniel Rubin, Charles E Kahn Jr, and Henning Müller. Comparing image search behaviour in the arrs goldminer search engine and a clinical pacs/ris. *Journal of biomedical informatics*, 56:57–64, 2015.

[DH15] Alba García Seco De Herrera. *Use Case Oriented Medical Visual Information Retrieval & System Evaluation*. PhD thesis, éditeur non identifié, 2015.

[DFAS+10] D Demner-Fushman, SK Antani, M Simpson, and Md M Rahman. Combining text and visual features for biomedical information retrieval. *Internal Technical Report, NIH*, 2010.

[DFKR+16] Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310, 2016.

[DDS+09] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009.

[DCL+19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[DMM19] Bhawna Dhruv, Neetu Mittal, and Megha Modi. Study of haralick's and glcm texture analysis on 3d medical images. *International Journal of Neuroscience*, 129(4):350–362, 2019.

[DAK10] B Dinakaran, J Annapurna, and Ch Aswani Kumar. Interactive image retrieval using text and image content. *Cybern Inf Tech*, 10:20–30, 2010.

[DWB+10] Bao H Do, Andrew Wu, Sandip Biswal, Aya Kamaya, and Daniel L Rubin. Informatics in radiology: radtf: a semantic search–enabled, natural language processor–generated radiology teaching file. *Radiographics*, 30(7):2039–2048, 2010.

[DCK+00] Molla S Donaldson, Janet M Corrigan, Linda T Kohn, et al. To err is human: building a safer health system, 2000.

[DLF+17] Yueqi Duan, Jiwen Lu, Jianjiang Feng, and Jie Zhou. Deep localized metric learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(10):2644–2656, 2017.

[FFK+13] Pawel Filipczuk, Thomas Fevens, Adam Krzyzak, and Roman Monczak. Computer-aided breast cancer diagnosis based on the analysis of cytological images of fine needle biopsies. *IEEE Trans. Medical Imaging*, 32(12):2169–2178, 2013.

[Fon20] Ruth Fong. *Understanding convolutional neural networks*. PhD thesis, University of Oxford, UK, 2020.

[FFY92] Christopher Fox, WB Frakes, and RB Yates. Lexical analysis and stoplist in information retrieval data structures and algorithms. *Prentice Hill*, 1992.

[Fra92] W Frakes. Stemming algorithms. In Norman O. Frederiksen and Harold Gulliksen, editors, *Information Retrieval: Data Structures and Algorithms*, pages 131–160. Prentice Hall, New York, New York, USA, 1992.

[G20] Keshav G. Multi-similarity loss. 2020. URL: https://kshavg.medium.com/multi-similarity-loss-for-deep-metric-learning-ad194691e2d3. Visited on 18/08/2022.

[GCR13] Nophar Geifman, Raphael Cohen, and Eitan Rubin. Redefining meaningful age groups in the context of disease. *Age*, 35(6):2357–2366, 2013.

[GAL+11] Payel Ghosh, Sameer Antani, L Rodney Long, and George R Thoma. Review of medical image retrieval systems and future directions. In *2011 24th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 1–6. IEEE, 2011.

[Gia18] Ivan Giangreco. *Database support for large-scale multimedia retrieval*. PhD thesis, University of Basel, 2018.

[GMK11] Yiannis Gkoufas, Anna Morou, and Theodore Kalamboukis. Suppl 1: combining textual and visual information for image retrieval in the medical domain. *The open medical informatics journal*, 5:50, 2011.

[Goa20] Keith Goatman. Computer vision for medical imaging applications. In 2020. Tutorial, ECCV 2020.

[GBC16a] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Autoencoders. In *Deep Learning*, chapter 14, pages 493–515. MIT Press, Cambridge, Massachusetts, 2016.

[GBC16b] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Convolutional networks. In *Deep Learning*, chapter 9, pages 321–359. MIT Press, Cambridge, Massachusetts, 2016.

[GBC16c] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Deep feedforward networks. In *Deep Learning*, chapter 6, pages 163–217. MIT Press, Cambridge, Massachusetts, 2016.

[GBC16d] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Introduction. In *Deep Learning*, chapter 1, pages 1–26. MIT Press, Cambridge, Massachusetts, 2016.

[GBC16e] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Machine learning basics. In *Deep Learning*, chapter 5, pages 95–151. MIT Press, Cambridge, Massachusetts, 2016.

[GBC16f] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. Optimization for training deep models. In *Deep Learning*, chapter 8, pages 267–309. MIT Press, Cambridge, Massachusetts, 2016.

[GLM+21] Mara Graziani, Thomas Lompech, Henning Müller, Adrien Depeursinge, and Vincent Andrearczyk. On the scale invariance in state of the art cnns trained on imagenet. *Machine Learning and Knowledge Extraction*, 3(2):374–391, 2021.

[GKK+02] Mark Oliver Gueld, Michael Kohnen, Daniel Keysers, Henning Schubert, Berthold B Wein, Joerg Bredno, and Thomas Martin Lehmann. Quality of dicom header information for image categorization. In *Medical imaging 2002: PACS and integrated medical information systems: design and evaluation*, volume 4685, pages 280–287. SPIE, 2002.

[GP19] Richard B Gunderman and Parth Patel. Perception's crucial role in radiology education. *Academic radiology*, 26(1):141–143, 2019.

[HSD73] Robert M Haralick, Karthikeyan Shanmugam, and Its' Hak Dinstein. Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6):610–621, 1973.

[HAA+20] Mehdi Hassan, Safdar Ali, Hani Alquhayz, and Khushbakht Safdar. Developing intelligent medical image modality classification system using deep transfer learning and lda. *Scientific reports*, 10(1):1–14, 2020.

[HZR+16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[HDG+07] Marti A Hearst, Anna Divoli, Harendra Guturu, Alex Ksikes, Preslav Nakov, Michael A Wooldridge, and Jerry Ye. Biotext search engine: beyond abstract search. *Bioinformatics*, 23(16):2196–2197, 2007.

[Her15] William R. Hersh. Information retrieval for healthcare. In Chandan K. Reddy and Charu C. Aggarwal, editors, *Healthcare Data Analytics*, pages 467–505. Chapman and Hall/CRC, 2015.

[HL15a] Johannes Hofmanninger and Georg Langs. Mapping visual features to semantic profiles for retrieval in medical imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 457–465, 2015.

[HL15b] Johannes Hofmanninger and Georg Langs. Mapping visual features to semantic profiles for retrieval in medical imaging. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 457–465. IEEE Computer Society, 2015.

[HSW89] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[HLM+17] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2261–2269. IEEE Computer Society, 2017.

[HPS+20] Shih-Cheng Huang, Anuj Pareek, Saeed Seyyedi, Imon Banerjee, and Matthew P Lungren. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ digital medicine*, 3(1):1–9, 2020.

[HLC12] Kyung Hoon Hwang, Haejun Lee, and Duckjoo Choi. Medical image retrieval: past and present. *Healthcare informatics research*, 18(1):3–9, 2012.

[ICD22] ICD-11. Icd-11 — Wikipedia, the free encyclopedia. 2022. URL: https://en.wikipedia.org/wiki/ICD-11. Visited on 17/02/2022.

[Ima12] ImageCLEF. Medical image classification and retrieval 2012. 2012. URL: https://www.imageclef.org/2012/medical. Visited on 04/03/2022.

[IRK+19] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 590–597. AAAI Press, 2019.

[ITM+18] Jason N Itri, Rafel R Tappouni, Rachel O McEachern, Arthur J Pesch, and Sohil H Patel. Fundamentals of diagnostic error in imaging. *Radiographics*, 38(6):1845–1865, 2018.

[JCB+20] Adam Jacobi, Michael Chung, Adam Bernheim, and Corey Eber. Portable chest x-ray in coronavirus disease-19 (covid-19): a pictorial review. *Clinical imaging*, 64:35–42, 2020.

[JCA+14] Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6):475, 2014.

[JXY+20] Shibo Jiang, Shuai Xia, Tianlei Ying, and Lu Lu. A novel coronavirus (2019-ncov) causing pneumonia-associated respiratory syndrome. *Cellular & molecular immunology*, 17(5):554–554, 2020.

[JPB+19] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):1–8, 2019.

[KJT07] Charles E Kahn Jr and Cheng Thao. Goldminer: a radiology image search engine. *American Journal of Roentgenology*, 188(6):1475–1478, 2007.

[KB19] Mahmut Kaya and Hasan Sakir Bilge. Deep metric learning: A survey. *Symmetry*, 11(9):1066, 2019.

[KRD+16] Brendan S Kelly, Louise A Rainford, Sarah P Darcy, Eoin C Kavanagh, and Rachel J Toomey. The development of expertise in radiology: in chest radiograph interpretation,"expert" search pattern may predate "expert"

levels of diagnostic accuracy for pneumothorax identification. *Radiology*, 280(1):252–260, 2016.

[Key23] Radiology Key. Projection x-ray imaging. 2023. URL: https://radiologykey.com/projection-x-ray-imaging/#c6_2. Visited on 17/02/2023.

[KKY+16] Kazuhiro Kitajima, Tomonori Kanda, Tomohiko Yamane, Tetsuya Tsujikawa, Hayato Kaida, Yukihisa Tamaki, Kozo Kuribayashi, et al. Present and future roles of fdg-pet/ct imaging in the management of lung cancer. *Japanese Journal of Radiology*, 34(6):387–399, 2016.

[Kit22] Dr Sean Kitson. Diagnostic imaging. 2022. URL: https://openmedscience.com/. Visited on 17/10/2022.

[KSH17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[KMDV+19] Philip J Kroth, Nancy Morioka-Douglas, Sharry Veres, Stewart Babbott, Sara Poplau, Fares Qeadan, Carolyn Parshall, Kathryne Corrigan, and Mark Linzer. Association of electronic health record design and use factors with clinician stress and burnout. *JAMA Network Open*, 2(8):e199609–e199609, 2019.

[LBH15] Yann LeCun, Yoshua Bengio, and Geoffrey E. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[LRL+19] Hongwei Li, Maximilian Reichert, Kanru Lin, et al. Differential Diagnosis for Pancreatic Cysts in CT Scans Using Densely-Connected Convolutional Networks. In *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC 2019)*, pages 2095–2098, Berlin, Germany. IEEE, July 2019.

[LGW+20] Qun Li, Xuhua Guan, Peng Wu, Xiaoye Wang, Lei Zhou, Yeqing Tong, Ruiqi Ren, Kathy SM Leung, Eric HY Lau, Jessica Y Wong, et al. Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. *New England journal of medicine*, 2020.

[LLC+13] Wei Li, B Li, P Cao, D Zhao, and J Yang. Combining text and content based image retrieval on medical resource database. In *ICMT 2013. Proceedings of the 3rd International Conference on Multimedia Technology*, pages 1771–1783, 2013.

[LZM+18] Zhongyu Li, Xiaofan Zhang, Henning Müller, and Shaoting Zhang. Large-scale retrieval for medical image analytics: A comprehensive review. *Medical Image Anal.*, 43:66–84, 2018.

[LTK16] Xinran Liu, Hamid R. Tizhoosh, and Jonathan Kofman. Generating binary tags for fast medical image retrieval based on convolutional nets and radon transform. In *2016 International Joint Conference on Neural Networks, IJCNN 2016, Vancouver, BC, Canada, July 24-29, 2016*, pages 2872–2878. IEEE, 2016.

[LJE+19] Yuan Liu, Ayush Jain, Clara Eng, David H. Way, Kang Lee, Peggy Bui, Kimberly Kanada, Guilherme de Oliveira Marinho, Jessica Gallegos, Sara Gabriele, Vishakha Gupta, Nalini Singh, Vivek Natarajan, Rainer Hofmann-Wellenhof, Gregory S. Corrado, Lily H. Peng, Dale R. Webster, Dennis Ai, Susan Huang, Yun Liu, R. Carter Dunn, and David Coz. A deep learning system for differential diagnosis of skin diseases. *CoRR*, abs/1909.05382, 2019.

[Low04] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 60(2):91–110, 2004.

[Mac22] Macmillan. Pet or pet-ct scan. 2022. URL: https://www.macmillan.org.uk/cancer-information-and-support/diagnostic-tests/pet-ct-scan. Visited on 7/10/2022.

[MS12] Tomas Majtner and David Svoboda. Extension of tamura texture features for 3d fluorescence microscopy. In *2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission*, pages 301–307. IEEE, 2012.

[MN19] Chris Manning and Pandu Nayak. Introduction to information retrieval. Lecture slides, 2019. Stanford University.

[MRS10] Christopher Manning, Prabhakar Raghavan, and Hinrich Schutze. Introduction to information retrieval. *Natural Language Engineering*, 16(1):100–103, 2010.

[MS20] Ashery Mbilinyi and Heiko Schuldt. Cross-modality medical image retrieval with deep features. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2632–2639. IEEE, 2020.

[MS21] Ashery Mbilinyi and Heiko Schuldt. Retrieving chest x-rays for differential diagnosis: a deep metric learning approach. In *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 1–4. IEEE, 2021.

[MS22] Ashery Mbilinyi and Heiko Schuldt. Cheres: a deep learning-based multifaceted system for similarity search of chest x-rays. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, pages 669–676, 2022.

[MSD+20] Edward R Melnick, Christine A Sinsky, Liselotte N Dyrbye, Mickey Trockel, Colin P West, Laurence Nedelec, and Tait Shanafelt. Association of perceived electronic health record usability with patient interactions and work-life integration among us physicians. *JAMA Network Open*, 3(6):e207374–e207374, 2020.

[Mic04] Wirth Michael. Texture analysis. Lecture slides, 2004. University of Guelph.

[MCC+13] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[MC17] Bhaskar Mitra and Nick Craswell. Neural models for information retrieval. *arXiv preprint arXiv:1705.01509*, 2017.

[MNZ+15] Hassan Mohamed, Abdelazim Negm, Mohamed Zahran, and Oliver C Saavedra. Assessment of artificial neural network for bathymetry estimation using high resolution satellite imagery in shallow lakes: case study el burullus lake. In *International water technology conference*, pages 12–14, 2015.

[Mül20] Henning Müller. Medical image retrieval: applications and resources. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 2–3, 2020.

[MHK+12] Henning Müller, Alba Garcia Seco de Herrera, Jayashree Kalpathy-Cramer, Dina Demner-Fushman, Sameer K. Antani, and Ivan Eggel. Overview of the imageclef 2012 medical image retrieval and classification tasks. In Pamela Forner, Jussi Karlgren, and Christa Womser-Hacker, editors, *CLEF 2012 Evaluation Labs and Workshop, Online Working Notes,*

*Rome, Italy, September 17-20, 2012*, volume 1178 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2012.

[MÜ17] Henning Müller and Devrim Ünay. Retrieval from and understanding of large-scale multi-modal medical datasets: A review. *IEEE Trans. Multim.*, 19(9):2093–2104, 2017.

[NGB17] Loris Nanni, Stefano Ghidoni, and Sheryl Brahnam. Handcrafted vs. non-handcrafted features for computer vision classification. *Pattern Recognit.*, 71:158–172, 2017.

[NLB10] Loris Nanni, Alessandra Lumini, and Sheryl Brahnam. Local binary patterns variants as texture descriptors for medical image analysis. *Artif. Intell. Medicine*, 49(2):117–125, 2010.

[NLL+22] Ha Q Nguyen, Khanh Lam, Linh T Le, Hieu H Pham, Dat Q Tran, Dung B Nguyen, Dung D Le, Chi M Pham, Hang TT Tong, Diep H Dinh, et al. Vindr-cxr: an open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9(1):1–7, 2022.

[BIB22a] National Institute of Biomedical Imaging and Bioengineering. Magnetic resonance imaging. 2022. URL: https://www.nibib.nih.gov/science-education/science-topics/magnetic-resonance-imaging-mri. Visited on 6/10/2022.

[BIB22b] National Institute of Biomedical Imaging and Bioengineering. Ultrasound. 2022. URL: https://www.nibib.nih.gov/science-education/science-topics/ultrasound. Visited on 6/10/2022.

[oMe07] National Library of Medicine. Pubmed user guide. 2007. URL: https://pubmed.ncbi.nlm.nih.gov/help/#help-stopwords. Visited on 01/04/2022.

[OPM02] Timo Ojala, Matti Pietikainen, and Topi Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7):971–987, 2002.

[OSM+17] John Edison Arevalo Ovalle, Thamar Solorio, Manuel Montes-y-Gómez, and Fabio A. González. Gated multimodal units for information fusion. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017.

[Özt20] Şaban Öztürk. Stacked auto-encoder based tagging with deep features for content-based medical image retrieval. *Expert Systems with Applications*, 161:113693, 2020.

[PSM14] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL, 2014.

[Pne22] Pneumonia. Pneumonia — Wikipedia, the free encyclopedia. 2022. URL: https://en.wikipedia.org/wiki/Pneumonia. Visited on 17/02/2022.

[Pok19] Sabina Pokhrel. How does computer understand images? 2019. URL: https://towardsdatascience.com/how-does-computer-understand-images-c1566d4537bf. Visited on 24/02/2022.

[PK11] BG Prasad and AN Krishna. Statistical texture feature-based retrieval and performance evaluation of ct brain images. In *2011 3rd International Conference on Electronics Computer Technology*, volume 2, pages 289–293. IEEE, 2011.

[RZK+19] Maithra Raghu, Chiyuan Zhang, Jon M. Kleinberg, and Samy Bengio. Transfusion: understanding transfer learning for medical imaging. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3342–3352, 2019.

[RIZ+17] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[RBH14] Yasser A. Reyad, Mohamed A. Berbar, and Muhammad Hussain. Comparison of statistical, lbp, and multi-resolution analysis features for breast mass classification. *J. Medical Systems*, 38(9):100, 2014.

[RHP+22] Sebastian Röhrich, Benedikt H Heidinger, Florian Prayer, Michael Weber, Markus Krenn, Rui Zhang, Julie Sufana, Jakob Scheithe, Incifer Kanbur, Aida Korajac, et al. Impact of a content-based image retrieval system on

the interpretation of chest cts of patients with diffuse parenchymal lung disease. *European Radiology*:1–8, 2022.

[Rud16] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[RDG+95] David E Rumelhart, Richard Durbin, Richard Golden, and Yves Chauvin. Backpropagation: the basic theory. *Backpropagation: Theory, architectures and applications*:1–34, 1995.

[SSG+19] NJ Sairamya, L Susmitha, S Thomas George, and MSP Subathra. Hybrid approach for classification of electroencephalographic signals using time–frequency images with wavelets and texture features. In *Intelligent Data Analysis for Biomedical Applications*, pages 253–273. Elsevier, 2019.

[SC12] Mark Sanderson and W. Bruce Croft. The history of information retrieval research. *Proc. IEEE*, 100(Centennial-Issue):1444–1451, 2012.

[SHZ+18] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: inverted residuals and linear bottlenecks. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 4510–4520. IEEE Computer Society, 2018.

[SKP15] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A Unified Embedding for Face Recognition and Clustering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, pages 815–823, Boston, MA, USA. IEEE, June 2015.

[SCN+22] Andrew B Sellergren, Christina Chen, Zaid Nabulsi, Yuanzhen Li, Aaron Maschinot, Aaron Sarna, Jenny Huang, Charles Lau, Sreenivasa Raju Kalidindi, Mozziyar Etemadi, et al. Simplified transfer learning for chest radiography models using less data. *Radiology*:212482, 2022.

[SCN+16] Amit Shah, Sailesh Conjeti, Nassir Navab, and Amin Katouzian. Deeply learnt hashing forests for content based image retrieval in prostate MR images. In Martin A. Styner and Elsa D. Angelini, editors, *Medical Imaging 2016: Image Processing, San Diego, California, USA, February 27, 2016*, volume 9784 of *SPIE Proceedings*, page 978414. SPIE, 2016.

[SUO+16] S Sharma, I Umar, L Ospina, D Wong, and Hamid R Tizhoosh. Stacked autoencoders for medical image search. In *International Symposium on Visual Computing*, pages 45–54. Springer, 2016.

[SKI+00] Junji Shiraishi, Shigehiko Katsuragawa, Junpei Ikezoe, Tsuneo Matsumoto, Takeshi Kobayashi, Ken-ichi Komatsu, Mitate Matsui, Hiroshi Fujita, Yoshie Kodera, and Kunio Doi. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American Journal of Roentgenology*, 174(1):71–74, 2000.

[SPC+20] Wilson Silva, Alexander Poellinger, Jaime S Cardoso, and Mauricio Reyes. Interpretability-guided content-based medical image retrieval. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 305–314. Springer, 2020.

[SZ14] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[SMT14] Hardeep Singh, Ashley ND Meyer, and Eric J Thomas. The frequency of diagnostic errors in outpatient care: estimations from three large observational studies involving us adult populations. *BMJ quality & safety*, 23(9):727–731, 2014.

[SZ03] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *9th IEEE International Conference on Computer Vision (ICCV 2003), 14-17 October 2003, Nice, France*, pages 1470–1477. IEEE Computer Society, 2003.

[Smi22] Robin Smithuis. Chest x-ray - lung disease: four-pattern approach. 2022. URL: https://radiologyassistant.nl/chest/chest-x-ray/lung-disease. Visited on 31/03/2022.

[SCE+10] Yang Song, Weidong Cai, Stefan Eberl, Michael J Fulham, and Dagan Feng. A content-based image retrieval framework for multi-modality lung images. In *2010 IEEE 23rd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 285–290. IEEE, 2010.

[SYS+17] Qinpei Sun, Yuanyuan Yang, Jianyong Sun, Zhiming Yang, and Jianguo Zhang. Using deep learning for content-based medical image retrieval. In *Medical Imaging 2017: Imaging Informatics for Healthcare, Research, and Applications*, volume 10138, page 1013812. International Society for Optics and Photonics, 2017.

[SIV+17] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In Satinder P. Singh and Shaul Markovitch, editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4278–4284. AAAI Press, 2017.

[TSG+16] Nima Tajbakhsh, Jae Y Shin, Suryakanth R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.

[TMY78] Hideyuki Tamura, Shunji Mori, and Takashi Yamawaki. Textural features corresponding to visual perception. *IEEE Transactions on Systems, man, and cybernetics*, 8(6):460–473, 1978.

[TZZ+12] Fabio Tavora, Y Zhang, M Zhang, L Li, M Ripple, D Fowler, and Allen Burke. Cardiomegaly is a common arrhythmogenic substrate in adult sudden cardiac deaths, and is associated with obesity. *Pathology*, 44(3):187–191, 2012.

[Tho23] Thoracickey. Basic principles in computer tomography (ct. 2023. URL: https://thoracickey.com/basic-principles-in-computed-tomography-ct/. Visited on 06/04/2023.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

[WGA+19] Stephen Waite, Arkadij Grigorian, Robert G Alexander, Stephen L Macknik, Marisa Carrasco, David J Heeger, and Susana Martinez-Conde. Analysis of perceptual expertise in radiology–current knowledge and a new perspective. *Frontiers in human neuroscience*, 13:213, 2019.

[WLQ+15] Yinan Wan, Fuhui Long, Lei Qu, Hang Xiao, Michael Hawrylycz, Eugene W Myers, and Hanchuan Peng. Blastneuron for automated comparison, retrieval and clustering of 3d neuron morphologies. *Neuroinformatics*, 13(4):487–499, 2015.

[WPL+17] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. Chestx-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3462–3471. IEEE Computer Society, 2017.

[WHH+19] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R. Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5022–5030. Computer Vision Foundation / IEEE, 2019.

[WTK+21] Yibo Wang, Amara Tariq, Fiza Khan, Judy Wawira Gichoya, Hari Trivedi, and Imon Banerjee. Query bot for retrieving patients clinical history: a covid-19 use-case. *Journal of biomedical informatics*, 123:103918, 2021.

[WHO22a] WHO. International classification of diseases 11th revision. 2022. URL: https://icd.who.int/en. Visited on 17/02/2022.

[WHO22b] WHO. Naming the coronavirus disease (covid-19) and the virus that causes it. 2022. URL: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it. Visited on 17/02/2022.

[WHO22c] WHO. Who coronavirus (covid-19) dashboard. 2022. URL: https://covid19.who.int/?adgroupsurvey={adgroupsurvey}&gclid=Cj0KCQiA1NebBhDDARIsAANiDD1QvxM0YYCYcnRIVDnisSOXk_J0dUmgdWWsvDgBi87d-QQArT2Ea6saAs3VEALw_wcB. Visited on 16/11/2022.

[TFE21] Wikipedia, The Free Encyclopedia. Spearman's rank correlation coefficient. 2021. URL: "https://en.wikipedia.org/wiki/Rank_correlation". Visited on 27/05/2021.

[TFE22a] Wikipedia, The Free Encyclopedia. Covid-19 pandemic. 2022. URL: https://en.wikipedia.org/wiki/COVID-19_pandemic. Visited on 16/11/2022.

[TFE22b] Wikipedia, The Free Encyclopedia. Differential diagnosis. 2022. URL: https://en.wikipedia.org/wiki/Differential_diagnosis#:~:text=In. Visited on 16/08/2022.

[TFE22c] Wikipedia, The Free Encyclopedia. Metric space. 2022. URL: https://en.wikipedia.org/wiki/Metric_space. Visited on 27/04/2022.

[TFE22d] Wikipedia, The Free Encyclopedia. Metric space. 2022. URL: https://en.wikipedia.org/wiki/Minkowski_distance. Visited on 04/04/2022.

[TFE22e] Wikipedia, The Free Encyclopedia. Pectral theorem. 2022. URL: https://en.wikipedia.org/wiki/Spectral_theorem. Visited on 29/07/2022.

[TFE22f] Wikipedia, The Free Encyclopedia. Similarity search. 2022. URL: https://en.wikipedia.org/wiki/Similarity_search. Visited on 27/04/2022.

[TFE23] Wikipedia, The Free Encyclopedia. Tf-idf. 2023. URL: https://en.wikipedia.org/wiki/Tf-idf. Visited on 18/02/2023.

[WKW+15] Guorong Wu, Minjeong Kim, Qian Wang, Brent C Munsell, and Dinggang Shen. Scalable high-performance image registration framework by unsupervised deep feature representations learning. *IEEE Transactions on Biomedical Engineering*, 63(7):1505–1516, 2015.

[XNH+18] Sun Xiaoming, Zhang Ning, Wu Haibin, Yu Xiaoyang, Wu Xue, and Yu Shuang. Medical image retrieval approach by texture features fusion based on hausdorff distance. *Mathematical Problems in Engineering*, 2018, 2018.

[XGF16] Junyuan Xie, Ross B. Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In Maria-Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 478–487. JMLR.org, 2016.

[XSC+10] MX Xin, ZX Shi, GB Cui, and HB Lu. Algorithm improvement of tamura texture features in content-based medical image retrieval. *Chinese Medical Equipment Journal*, 31:32–35, 2010.

[XY16] Fuyong Xing and Lin Yang. Robust nucleus/cell detection and segmentation in digital pathology and microscopy images: a comprehensive review. *IEEE reviews in biomedical engineering*, 9:234–263, 2016.

[XMK08] Songhua Xu, James McCusker, and Michael Krauthammer. Yale image finder (yif): a new search engine for retrieving biomedical images. *Bioinformatics*, 24(17):1968–1970, 2008.

[YC08] Hong Yu and Yong-gang Cao. Automatically extracting information needs from ad hoc clinical questions. In *AMIA annual symposium proceedings*, volume 2008, page 96. American Medical Informatics Association, 2008.

[YHL+21] Yang Yu, Peng Hu, Jie Lin, and Pavitra Krishnaswamy. Multimodal multitask deep learning for x-ray image retrieval. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 603–613. Springer, 2021.

[ZF14] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, volume 8689 of *Lecture Notes in Computer Science*, pages 818–833. Springer, 2014.

[ZZ09] Ethan Zhang and Yi Zhang. *Average precision.* In *Encyclopedia of Database Systems.* LING LIU and M. TAMER ÖZSU, editors. Springer US, Boston, MA, 2009, pages 192–193. ISBN: 978-0-387-39940-9.

[ZSC+17] Fan Zhang, Yang Song, Weidong Cai, Adrien Depeursinge, and Henning Müller. Text-and content-based medical image retrieval in the visceral retrieval benchmark. 2017.

[ZLD+15] Xiaofan Zhang, Wei Liu, Murat Dundar, Sunil Badve, and Shaoting Zhang. Towards large-scale histopathological image analysis: hashing-based image retrieval. *IEEE Trans. Medical Imaging*, 34(2):496–506, 2015.

[ZFL+12] Jianlong Zhou, Chang Feng, Xiaoming Liu, and J Tang. A texture features based medical image retrieval system for breast cancer. In *2012 7th International Conference on Computing and Convergence Technology (ICCCT)*, pages 1010–1015. IEEE, 2012.

[ZGD+20] S. Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S. Duncan, Bram van Ginneken, Anant Madabhushi, Jerry L. Prince, Daniel Rueckert, and Ronald M. Summers. A review of deep learning in medical imaging: image traits, technologytrends, case studies with progress highlights, and future promises. *CoRR*, abs/2008.09104, 2020.

[ZZW+20] Na Zhu, Dingyu Zhang, Wenling Wang, Xingwang Li, Bo Yang, Jingdong Song, Xiang Zhao, Baoying Huang, Weifeng Shi, Roujian Lu, et al. A novel coronavirus from patients with pneumonia in china, 2019. *New England journal of medicine*, 2020.

[ZVS+18] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710, 2018.