ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ
ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

School of Information Sciences and Technology
Department of Informatics
Athens, Greece

Bachelor Thesis
in
Informatics

Enhanced Biomedical Image Tagging

Anna Chatzipapadopoulou

Supervisors: Prof. Ion Androutsopoulos
Department of Informatics
Athens University of Economics and Business

Assistant Prof. John Pavlopoulos
Department of Informatics
Athens University of Economics and Business

September 2024

Anna Chatzipapadopoulou

Enhanced Biomedical Image Tagging

September 2024

Supervisors: Prof. Ion Androutsopoulos, Assistant Prof. John Pavlopoulos

Athens University of Economics and Business

School of Information Sciences and Technology

Department of Informatics

Information Processing Laboratory

Athens, Greece

Abstract

Medical image classification is a complex and fascinating task that has significantly benefited from advancements in deep learning techniques. This thesis focuses on the specific challenge of multi-class multi-label medical image classification, also known as medical image tagging. The objective is to accurately assign relevant medical terms (concepts/tags) to images, which describe potential findings and ultimately assist clinicians in the diagnostic process. Medical image tagging may involve analyzing images from various modalities, such as X-rays, MRIs, CT scans, and ultrasonographies, to identify and categorize different pathological conditions. To address the task of medical image tagging, we develop and evaluate various deep learning models that encode the images and combine them with both classification-based and retrieval-based approaches.

Ultimately, our goal is to develop a robust and reliable system for medical image tagging that not only provides clinicians with supportive information but also assists them in diagnosing diseases more accurately and efficiently.

Περίληψη

Η ταξινόμηση ιατρικών εικόνων αποτελεί μια απαιτητική και πολυδιάστατη διαδικασία, που έχει επωφεληθεί σημαντικά από τις εξελίξεις στον τομέα της βαθιάς μάθησης. Στην παρούσα εργασία, επικεντρωνόμαστε στην πρόκληση της ταξινόμησης ιατρικών εικόνων που ανήκουν ταυτόχρονα σε πολλές κατηγορίες, η οποία είναι γνωστή και ως κατηγοριοποίηση ιατρικών εικόνων (biomedical image tagging). Στόχος είναι η ακριβής συσχέτιση των ιατρικών εικόνων με τις κατάλληλες ετικέτες (ιατρικούς όρους), οι οποίες περιγράφουν πιθανά ευρήματα και παθολογίες, υποστηρίζοντας έτσι τους κλινικούς ιατρούς στη διαδικασία της διάγνωσης.

Προκειμένου να επιτευχθεί αυτός ο στόχος, αναπτύσσουμε και αξιολογούμε προηγμένα μοντέλα βαθιάς μάθησης που κωδικοποιούν τις εικόνες και εφαρμόζουμε τόσο τεχνικές κατηγοριοποίησης όσο και τεχνικές ανάκτησης πληροφοριών. Οι τεχνικές κατηγοριοποίησης εκπαιδεύουν τα μοντέλα βαθιάς μάθησης ώστε να προβλέπουν με αυτόματο τρόπο τις κατάλληλες ετικέτες από τις εικόνες, αξιοποιώντας επισημειωμένα σύνολα δεδομένων. Αντίθετα, οι τεχνικές ανάκτησης συγκρίνουν νέες εικόνες με ήδη επισημασμένες εικόνες στη βάση δεδομένων, για να εντοπίσουν τις πιο παρόμοιες και να συναγάγουν τις ετικέτες βάσει αυτής της σύγκρισης.

Ο απώτερος στόχος είναι η δημιουργία ενός αξιόπιστου και αποτελεσματικού συστήματος ταξινόμησης ιατρικών εικόνων, το οποίο όχι μόνο θα βελτιώνει την ακρίβεια και την αποτελεσματικότητα της διάγνωσης, αλλά θα παρέχει και πολύτιμες πληροφορίες στους κλινικούς ιατρούς, βοηθώντας τους στη λήψη αποφάσεων.

Acknowledgements

I would like to express my sincere appreciation to my supervisors, Ion Androutsopoulos and John Pavlopoulos, for granting me the opportunity to collaborate with them and for their steadfast support throughout the development of this thesis. Their persistent guidance and encouragement greatly expanded my research and academic perspectives.

I am also deeply thankful to our biomedical team members, PhD candidates G. Moschovis and F. Charalampakos, along with master's students P. Kaliosis and M. Samprovalaki. Their insightful offline discussions and guidance on both theoretical and practical matters were invaluable throughout the duration of this project.

I also wish to express my deepest appreciation to my family and friends for their unwavering support and encouragement. Their love, patience, and understanding have been my foundation throughout this journey.

Contents

1 Introduction

Medical imaging plays a pivotal role in modern healthcare, providing critical insights that aid in the diagnosis, monitoring, and treatment of various medical conditions. With the advent of advanced imaging technologies such as X-rays, MRIs, PET/CT scans, and ultrasounds, the volume of medical images generated has increased exponentially [Naj22]. This surge in data necessitates the development of efficient and accurate methods for interpreting and managing medical images.

This thesis primarily focuses on large-scale multi-class multi-label medical image classification, also known as medical image tagging, which is illustrated in Figure 1.1. This task involves assigning *multiple* relevant medical terms (tags) to an image, which describe potential findings, anatomical structures, or pathological conditions present in the image. Such terms can significantly aid clinicians by providing an initial diagnostic report or supporting information, thereby enhancing the diagnostic process and improving patient outcomes. Traditionally, medical image interpretation heavily relies on the expertise of radiologists and other medical professionals. However, manual analysis is time-consuming, subject to human error, and often impractical, given the sheer volume of medical exams that need to be reviewed. Consequently, there is a growing need for automated systems that can assist in the accurate and efficient interpretation of medical images.

In this thesis, we explore several deep learning architectures to encode medical images effectively. We integrate these architectures with both classification-based and retrieval-based methods to enhance the accuracy and efficiency of image tagging. Classification-based methods involve training deep learning models to predict a set of medical concepts (tags) directly from the images, while retrieval-based methods compare query images to a database of labeled images to infer the most relevant tags. Additionally, we incorporate advanced techniques such as data augmentation, transfer learning [Zhu+20], and ensemble methods to further improve model performance. Data augmentation helps create a more diverse training set, enhancing the model's generalization capability. Transfer learning allows us to leverage pre-trained models on large datasets, facilitating more efficient training on medical images. Finally, ensemble methods combine predictions from multiple models to reduce variability and increase accuracy.

Part of this thesis concerns the participation of the AUEB NLP Group in the 2024 ImageCLEFmedical campaign [Sam+24], specifically in the Caption task [Ion+24; Rüc+24a]. Each year, participants from around the world are invited to develop and test innovative

algorithms capable of interpreting and classifying complex medical images, thereby contributing to the improvement of automated medical image analysis systems. The task is divided into two sub-tasks: Concept Detection and Caption Prediction. The former focuses on medical image tagging methods, while the latter concerns diagnostic captioning systems. The author's main contribution was to the Concept Detection sub-task, which is the primary focus of this thesis. Building on the group's previous successful participations [KPA19; KPA20; Cha+21; Cha+22], we achieved commendable results in this year's competition. We ranked 2^{nd} in the Concept Detection (tagging) task and 4^{th} in the Caption Prediction (captioning) task among 9 and 11 research groups respectively [Sam+24]. We will discuss the tagging systems, present our results, and highlight additional work conducted.

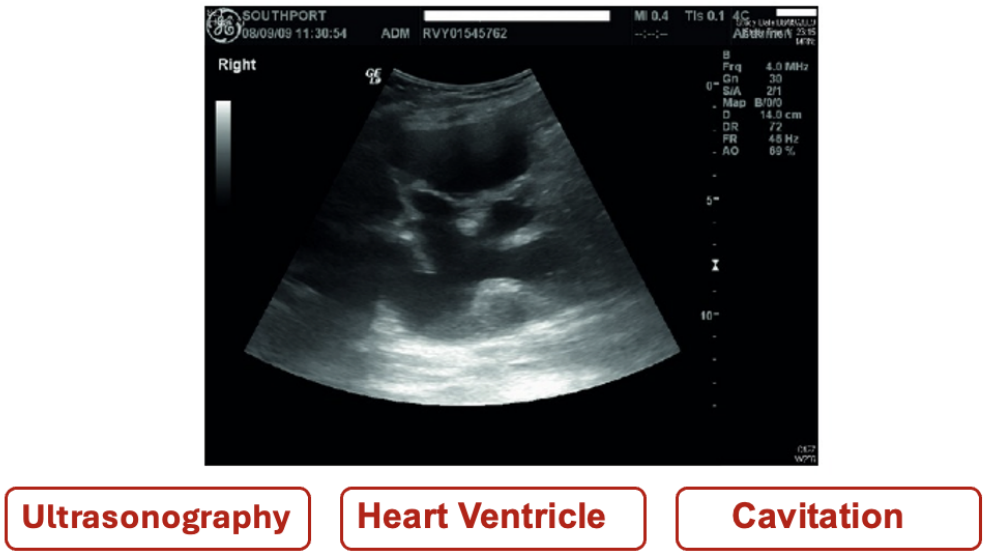


Fig. 1.1: This image illustrates medical image tagging, where key features from an ultrasound are labeled, such as "Ultrasonography," "Heart Ventricle," and "Cavitation". CC BY [Magdas et al.(2021)]

1.1 Motivation and Problem Statement

This study is motivated by the need to improve biomedical image tagging, which is essential for enhancing the accessibility and precision of AI-generated interpretations of medical information. Current tagging methods have room for improvement in accuracy, with the potential to enhance diagnostic processes and support smoother healthcare workflows. By developing more innovative and effective tagging techniques, this research aims to create systems that are both accurate and efficient, capable of managing the complexity of biomedical images with minimal manual input. The ultimate goal is to contribute to biomedical image processing by providing solutions that improve diagnostic speed, patient outcomes, and healthcare processes in both clinical and research settings.

1.2 Thesis Structure

Chapter 1: Introduction

This chapter introduces the importance of medical image classification in healthcare, discussing how advancements in imaging technologies and the sheer volume of medical images have created the need for automated tagging systems. It outlines the research motivation, problem statement, and the overall goals of this thesis, which focuses on improving multi-label classification for medical images to assist in diagnosis and image retrieval.

Chapter 2: Background and Related Work

This chapter provides an overview of existing methods and technologies in the field of image classification, with a focus on medical images. It discusses traditional methods, machine learning approaches, and deep learning techniques. Related work in both classification-based and retrieval-based techniques is reviewed, along with key challenges in multi-label medical image tagging.

Chapter 3: Implemented methods and Systems

This chapter describes the models and systems developed during the thesis. It focuses on deep learning-based methods, particularly Convolutional Neural Network (CNN) encoders combined with Feed-Forward Neural Networks (FFNN), and retrieval-based methods using k-nearest neighbors (k-NN). Details are provided on the architecture, design decisions, and variations of these methods, along with descriptions of the implemented algorithms.

Chapter 4: Data

In this chapter, the datasets used in the study are presented, including the ImageCLEFmedical 2024 dataset. The chapter covers the dataset's structure, characteristics, and the preprocessing steps applied to prepare the data for training and testing. An exploratory analysis is also provided to highlight key trends and imbalances in the dataset, such as the distribution of medical concepts and labels.

Chapter 5: Experiments and Results

This chapter presents the experimental setup and the results of applying the developed models. It discusses the training process, the evaluation metrics used (such as F_1 score), and the performance of various models. Detailed results are provided for both classification-based and retrieval-based approaches, comparing them in terms of their F_1 scores. The chapter also explores the impact of data augmentation and transfer learning on model performance.

Chapter 6: Conclusions and Future Work

This final chapter summarizes the key findings and contributions of the thesis. It also discusses the limitations of the current approaches and outlines potential directions for future research, including further improvements in multi-label classification and retrieval-based methods.

2 Background and Related Work

Image classification is a fundamental task in the field of computer vision. It involves assigning one or more predefined categories (labels) to an image based on its visual content. This requires analyzing the image's patterns to determine the appropriate classes from a predefined set of categories. The purpose of image classification is to organize and make sense of large volumes of image data, enabling efficient search, retrieval, and analysis. Modern advancements, particularly in Deep Learning (DL) and CNNs, have significantly improved the accuracy of image classification, making it an indispensable tool in numerous applications.

Image classification is particularly important in medical imaging, where vast amounts of data need to be processed and analyzed to identify critical features or detect diseases. From automated diagnostic tools to decision support systems, image classification plays a critical role in advancing healthcare technologies. Furthermore, medical image taggers can be particularly important in diagnostic captioning as well, when guided decoding is employed to steer a Language Model (LM) to produce more medically accurate reports. Distance from Median Maximum Cosine Similarity (DMMCS) [Kal+24] is a recently published framework describing the process of controlled generation of Vision-Language Models based on predicted concepts. It combines the original LM score with the DMMCS penalty to get the final generation probabilities of the diagnostic reports.

2.1 Background

Image classification has undergone significant advancements over the years, driven by both theoretical developments and technological innovations. From early manual techniques to today's automated systems, the field continues to evolve, addressing increasingly complex tasks with greater accuracy. These advancements have broad applications across various domains, including healthcare.

2.1.1 General Methods

There are two main ways to perform image classification:

- **Manual Classification** involves human experts who inspect and label images based on their knowledge and experience. While straightforward, this method can be labor-intensive and inconsistent due to human error and subjectivity. It is typically used when the number of images is small or when expert interpretation is required. For instance, in rare disease diagnostics, human expertise is crucial in identifying subtle anomalies. Manual classification tends to offer greater accuracy in cases that require expert interpretation, though it is less scalable and more time-consuming.

- **Automated Classification** leverages computational algorithms to categorize images without human intervention. This approach is scalable, allowing the handling of large datasets much more efficiently. Automated methods typically involve Machine Learning (ML) and DL techniques. These approaches have significantly advanced the field of medical imaging, where the automation of image labeling and analysis is critical for improving diagnostic workflows and reducing clinicians' workload. While automated classification is highly efficient and scalable, it may not always match the precision of human experts in specialized cases.

2.1.2 Progress in ML-Based Image Classification

The rapid advances in image classification, particularly in the domain of medical imaging, have led to significant developments. Early methods relied on feature extraction techniques such as Scale-Invariant Feature Transform (SIFT)[Low04] and HOG (Histogram of Oriented Gradients) [DT05], but these approaches had limited success when applied to complex medical images. The introduction of CNNs, as popularized by Krizhevsky et al. [KSH12] with the AlexNet architecture, marked a turning point in image classification. This advancement opened new possibilities for automated feature extraction, allowing networks to learn image features hierarchically. Since then, architectures such as ResNet [He+16] and EfficientNet [TL19] have pushed the boundaries in classification accuracy.

In the context of biomedical imaging, CNNs have been instrumental in automating the detection of diseases in medical images. For instance, they have shown remarkable success in detecting tumors from mammograms, identifying diabetic retinopathy from retinal scans, and classifying lung diseases from chest X-rays [Kas+24; BB22; Ift+24]. These breakthroughs demonstrate the significant impact that deep learning can have in clinical diagnostics, improving both speed and accuracy while reducing human error.

2.2 Image Classification

Image Classification problems can be distinguished into the following cases:

- **Binary Classification:** Binary classification involves categorizing images into one of two classes. This is the simplest form of image classification where the goal is to decide between two distinct categories. For example, in medical imaging, binary classification can be used to determine whether an X-ray image indicates the presence or absence of a disease. Another common application in binary classification tasks is, when the system must decide between two possible outcomes, such as determining whether a car is present in a given scenario or not. The nature of this task makes it relatively straightforward to implement, but it requires high accuracy, especially in critical applications like medical diagnostics [Lit+17].

- **Multi-class classification:** Multi-class classification extends the concept of binary classification to scenarios where there are more than two classes. In this case, each image is categorized into one of several possible categories. For example, a system might classify images of animals into different species. This type of classification is more complex due to the higher number of categories, each requiring a distinct set of features for accurate identification. In the medical field, multi-class classification could involve categorizing types of tumors or identifying different diseases from medical scans [Lee+20].

- **Multi-class, Multi-label classification (MCML):** Unlike multi-class classification where an image belongs to one category only, in multi-label classification, each image can belong to multiple categories simultaneously. This is particularly useful in scenarios where objects or certain features in an image are not mutually exclusive. For instance, a single image might contain both ‘people‘ and ‘vehicles‘, and it should be tagged with both labels. In medical imaging, an X-ray might show signs of multiple conditions, such as both pneumonia and a broken rib. MCML classification requires algorithms capable of handling the complexity of assigning multiple labels to a single instance, ensuring that all relevant features are recognized and tagged appropriately [Irv+19].

- **Hierarchical classification:** Hierarchical classification involves organizing categories into a hierarchy and classifying images at multiple levels of this hierarchy. This approach is particularly beneficial when dealing with complex datasets where categories can be naturally grouped into subcategories. For example, an image classification system might first determine if an image is of an animal or a plant. If it is an animal, it might then classify it as a mammal, bird, reptile, etc., and further down classify a mammal as a dog, cat, horse, etc. In the context of medical image tagging, hierarchical classification can be very valuable. For instance, a medical image classification system, as proposed by Zhou et al. [ZLZ20], might first categorize an image into broad groups, such as normal or abnormal tissue. Subsequently, abnormal images could be classified by their imaging modality (e.g., X-ray, ultrasonography,

MRI, or CT scan), followed by a finer classification of modality-specific features, such as the body part depicted in X-ray images (e.g., chest, spine, or extremities).

2.3 Techniques Used in Automated Image Classification

2.3.1 ML-based techniques

Supervised ML-based techniques use algorithms to learn patterns from labeled training data and classify new images. Key ML techniques include decision trees, k-nearest neighbors (k-NN), and random forests. These methods require feature extraction, where specific characteristics of the images are identified and used for classification.

2.3.2 DL-based techniques

DL, a subset of ML, employs neural networks with many layers (deep neural networks) to learn features and classify images. CNNs, a kind of DL, are particularly effective for image classification tasks. They automatically learn hierarchical features from images—ranging from low-level features like edges to high-level features like shapes and objects. CNNs have become the foundation for many state-of-the-art image classification systems due to their ability to learn from large datasets, such as ImageNet [Rus+15], and generalize well to unseen data.

Transfer Learning (TL) is often used to address the lack of large annotated datasets, especially in medical image classification. In a TL setting, models pre-trained on large datasets, such as ImageNet [Rus+15], are fine-tuned on smaller, domain-specific datasets (e.g., medical images). This approach allows for efficient training and better performance, even with limited labeled data, as the model has already acquired general knowledge that can be adapted to the task at hand. TL has been particularly successful in medical applications, where it helps address the challenge of limited data, speeding up model development and improving accuracy [Taj+20].

Following the success of Transformers [Vas+17a] in Natural Language Processing, with BERT-based encoders [Dev+19] being vastly used for feature extraction from text, Vision Transformers (ViT) [Dos+21] are also becoming popular for encoding images into dense numeric representations, especially for tasks requiring global feature understanding. However, CNNs have demonstrated superior performance compared to a ViT in semantic segmentation and generic visual classification [AMT22], as well as in encoding medical images for concept detection [MF22; Mos22]. Generative Adversarial Networks (GANs) have

also been explored to augment data, enabling better generalization for medical imaging tasks where data scarcity is an issue.

2.4 Architectures Used in Image Classification

The development of deep learning architectures has revolutionized image classification. Here, we delve into some of the most widely used architectures - Multi-Layer Perceptrons (MLPs), CNNs, and ViTs - by describing how they work and discussing their relevance to image classification, and particularly medical imaging.

2.4.1 MLPs

MLPs [Ros57; Ros62] are one of the earliest forms of neural networks, consisting of multiple layers of neurons connected in a 'fully connected manner': each neuron in one layer is connected to every neuron in the subsequent layer. An MLP generally includes three main components: an input layer, one or more hidden layers, and an output layer. The input layer receives the raw data (such as pixel values of an image), while the hidden layers process this input through weighted connections and apply non-linear activation functions like ReLU (Rectified Linear Unit). Lastly, the output layer produces the final predictions.

The learning process in MLPs involves backpropagation [RHW86], where the network's prediction is compared with the true label, and the resulting error is used to update the weights to minimize the loss. Over time, the model improves its ability to classify the input data correctly. However, MLPs have several limitations when it comes to image classification. One significant issue is that they treat the input as a one-dimensional vector, flattening the image and thereby losing crucial spatial information. Moreover, due to their fully connected nature, they require a large number of parameters, which makes them computationally expensive and prone to overfitting, particularly with high-dimensional data like images.

In practice, MLPs are rarely used as encoders for image classification tasks. They are often paired with other architectures that can better handle the spatial structure of image data. The inefficiency of MLPs in this domain led to the development of more specialized architectures, such as CNNs, which are designed to exploit the spatial locality of images.

2.4.2 CNNs

CNNs [LeC+89] represent a major leap forward in image classification by addressing the shortcomings of MLPs. CNNs are designed to work with grid-like data, such as images, by

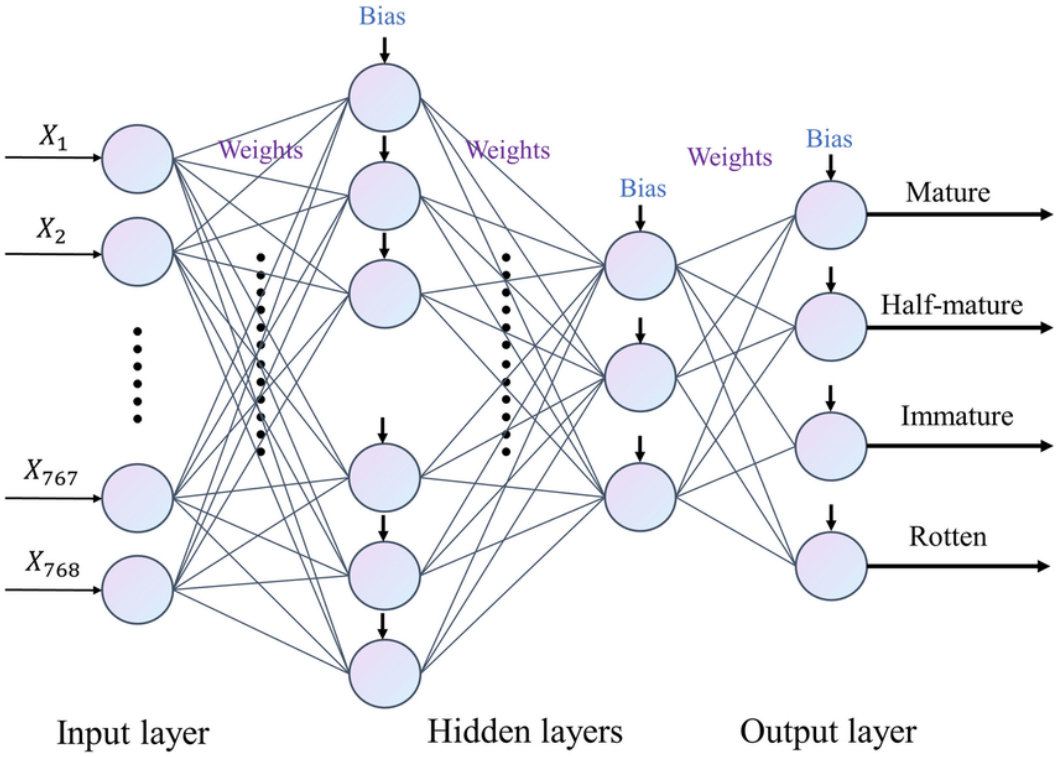


Fig. 2.1: Architecture of an MLP used for classifying strawberries into different appearance quality categories, such as mature, half-mature, immature, and rotten. The network consists of an input layer, multiple hidden layers, and an output layer. Each node in the hidden and output layers is fully connected to the nodes in the previous layer through learned weights and biases. The network takes input features (e.g., pixel values), processes them through hidden layers, and outputs predictions for different categories such as, in this case, "Mature," "Half-mature," "Immature," and "Rotten." Figure taken from [ZWL22]

preserving spatial hierarchies. The primary building block of CNNs is the convolutional layer, which applies filters (kernels) to the input image. These filters slide over the image, performing a dot product between the filter and the local regions of the image, resulting in a feature map that highlights specific patterns, such as edges or textures. This allows CNNs to detect local patterns and combine them into more complex structures as the network deepens. Another critical component of CNNs is the pooling layer, which reduces the dimensionality of the feature maps by summarizing local regions. A common pooling operation is max pooling, which takes the maximum value from each region of the feature map, helping the network retain the most important features while reducing computational load. After several layers of convolution and pooling, the feature maps are flattened and passed through fully connected layers, similar to those in an MLP, to produce the final classification output. CNNs are highly effective for image classification because they use local receptive fields, meaning that each neuron only processes a small part of the image. This enables CNNs to capture local features like edges and textures in early layers and combine them into more abstract features in deeper layers. Additionally, CNNs employ parameter sharing, where the same filters are used across the entire image, drastically reducing the number of parameters compared to fully connected networks. This not only

makes CNNs computationally more efficient than MLPs, but also allows them to generalize better to unseen data. In the field of medical imaging, CNNs have been used to detect abnormalities and diseases in various types of scans.

Some popular CNN architectures used in image classification include:

- **LeNet**: One of the earliest CNNs, originally developed for digit recognition but foundational for later developments [LeC+98].
- **AlexNet**: The architecture that won the 2012 ImageNet competition, showcasing the potential of deep CNNs [KSH12].
- **ResNet**: An innovative architecture with residual blocks that enables the training of much deeper networks [He+16].
- **EfficientNet**: A family of models that use a compound scaling method to balance depth, width, and resolution, offering state-of-the-art performance with fewer parameters [TL19].

2.4.3 Vision Transformers (ViTs)

Vision Transformers (ViTs) [DBK+21] are a relatively recent innovation in image classification, borrowing the Transformer architecture originally developed for natural language processing (NLP). Unlike CNNs, which focus on local features through convolutional operations, ViTs treat an image as a sequence of patches, similar to how transformers process sequences of words in text. An image is divided into fixed-size patches (for example, 16x16 pixels), each of which is flattened into a one-dimensional vector. These vectors are then embedded into a higher-dimensional space and processed by transformer layers.

One of the key mechanisms in ViTs is multi-head self-attention [Vas+17b], which allows each patch to attend to every other patch in the image. This enables the model to capture long-range dependencies between different parts of the image. For example, in a medical image, a ViT could recognize that distant patches correspond to different parts of an organ or structure, allowing for a more holistic understanding of the image. To retain information about the original spatial arrangement of the patches, position embeddings are added to the patch embeddings. The final classification in a ViT is typically done by using a special class token, which summarizes the information from all patches. This class token is passed through a fully connected layer to produce the final classification result. ViTs excel at capturing global relationships across an image, making them particularly powerful for tasks where understanding long-range dependencies is critical. This contrasts with CNNs, which primarily focus on local features. However, one challenge with ViTs is their reliance

on large amounts of training data to perform well. Without sufficient data, they may struggle to generalize, particularly in comparison to CNNs. Despite this, ViTs have shown promising results in medical imaging because they can capture relationships between different parts of an image, even if those parts are far apart. This makes them particularly useful for detecting subtle patterns or conditions that span across different regions of an image. While CNNs are also capable of learning complex patterns, they do so by first focusing on small, local regions and gradually combining these features as the network deepens. ViTs, on the other hand, process the image as a whole from the beginning, allowing them to recognize broader patterns and interactions between different areas right away, which can be especially beneficial in tasks where the connections between distant areas of the image are important for accurate diagnosis.

2.5 Medical Image Classification

Medical image classification is a specialized subfield of image classification that focuses on diagnosing diseases, identifying abnormalities, and aiding clinical decision-making through the analysis of medical images. Unlike generic image classification, which benefits from large and diverse datasets, medical image classification is constrained by data availability, regulatory standards, and the need for high interpretability. This section reviews the unique challenges and methodologies used in medical image classification, from traditional approaches to cutting-edge techniques.

With the rapid advancement of deep learning, particularly CNNs, there has been a transformative shift in how medical images are processed and analyzed. CNNs, which have become the cornerstone of medical image classification, have been successfully applied to a variety of diagnostic tasks, ranging from detecting pneumonia in chest X-rays to segmenting brain tumors in MRI scans [Rah+20; JS22].

One of the most notable examples of CNNs in medical imaging is CheXNet, proposed by Rajpurkar et al. [Raj+17], a deep learning model designed to detect pneumonia from chest X-rays. CheXNet was trained using transfer learning, leveraging a pre-trained model on the ImageNet dataset [KSH12] and fine-tuning it on medical images. This approach is common in medical imaging due to the scarcity of large, labeled datasets. Models that have been pre-trained on large generic datasets can be adapted to medical applications with significantly smaller amounts of domain-specific data. CheXNet, which was based on DenseNet [Hua+17], provided a foundation for further experimentation and significant performance improvements in medical imaging tasks. Several works, including [KPA19; KPA20; Kar+20; Cha+21; Cha+22], extended CheXNet by incorporating various CNN architectures such as EfficientNet [TL19], VGG [LD15], and ResNet [He+16]. In addition

to CNNs, they also explored the impact of Transformer-based architectures, like ViT, for image encoding.

These models were further optimized using ensemble methods, combining multiple instances of these architectures to improve performance. Typically, a FFNN was used to predict the presence of specific concepts in an image, where concept assignment depended on whether the predicted probability exceeded a set threshold. Several of these approaches secured top positions in the ImageCLEFmed concept detection task, achieving first place in 2019 [Pel+19], 2020 [Pel+20], 2021 [PBG+21], and 2022 [RBG+22].

Regarding biomedical image retrieval, initial approaches with simpler non-neural methods were also successful. For instance, Valavanis and Stathopoulos [VS17] achieved first place in the 2017 ImageCLEFmed [Eic+17] concept detection task by utilizing traditional retrieval techniques like Bag of Colours (BoC) and Bag of Visual Words (BoVW).

In most recent years, neural retrieval methods have dominated the ImageCLEFmed competition. In 2019, [KPA19] secured third place by leveraging a DenseNet-121 [Hua+17] image encoder pre-trained on ImageNet [Den+09], paired with the k-NN algorithm [CH67] to find the most similar images. The tags were then assigned based on a majority voting strategy. Further approaches explored various encoding architectures like ResNet[He+16], alongside different voting mechanisms, and also delivered competitive results. In 2023, AUEB's NLP group achieved first place in the ImageCLEFmed concept detection task with their top-performing model. Their winning approach [Kal+23] was a union ensemble that combined three instances of their CNN+FFNN system, utilizing different encoding backbones for each instance: EfficientNetB0 [TL19], EfficientNetB0v2, and DenseNet121 [Hua+17].

In 2021, Charalampakos et al. [Cha+21] won the competition with an ensemble of several variations of 1-NN classifiers, experimenting with different CNN image encoders such as EfficientNet [TL19], DenseNet [Hua+17], and ResNet [He+16]. Meanwhile, AUEB's NLP group investigated a weighted k-NN technique, and in 2022, their ensemble—combining two CNN+FFNN models and a CNN+wKNN architecture—placed fifth in the ImageCLEFmed concept detection task [RBG+22].

In addition to performance improvements, interpretability remains a critical aspect of medical image classification. Techniques like Grad-CAM [Sel+20] have been integrated into CNN-based systems to provide heatmaps that highlight the regions of an image contributing most to the model's decision. This is particularly important in medical applications, where clinicians might require an understanding of the model's reasoning to trust automated predictions.

Beyond 2D images, 3D CNNs have also emerged as a powerful tool for analyzing volumetric medical data, such as MRIs and CT scans. These models capture spatial information in three dimensions, making them particularly effective for tasks like brain tumor segmentation and heart structure analysis. For instance, Milletari et al. [MNA16] introduced V-Net, a 3D CNN architecture that uses a Dice loss function to improve performance on volumetric medical image segmentation tasks. Similarly, Chen et al. [Che+19] applied 3D CNNs for automated brain tumor segmentation, showing that 3D convolutions are essential for accurate medical imaging analysis in certain applications.

Generative models, particularly Generative Adversarial Networks (GANs), have also found applications in medical image classification. GANs are used for data augmentation, generating synthetic images to enhance training datasets and improve model performance. Shin et al. [Shi+18] demonstrated how GANs can be used to augment medical image datasets, leading to better classification results. More recently, diffusion models have emerged as the state-of-the-art (SoTA) in image generation, surpassing GANs in quality and diversity of generated images [DN21]. These models have shown remarkable success in generating high-fidelity medical images, further advancing data augmentation and model training in the field [Kha+22].

3 Implemented methods and Systems

This chapter presents the methods and systems that were implemented throughout the course of this thesis. Specifically, it presents four distinct methodologies: the combination of a CNN encoder with FFNNs, retrieval-based approaches, the integration of an image encoder with k-NN retrieval, and an innovative variation of the k-NN algorithm.

3.1 CNN encoder + FFNN

This system is based on previous work of the AUEB NLP Group [Cha+21; KPA19; Kal+23]. The system utilizes a CNN as its backbone for feature extraction, followed by a FFNN for classification, as seen in Figure 3.1. To generate an image embedding, we extract feature maps from the last convolutional layer of the CNN and condense them into a feature vector using the Generalized-Mean (GeM) pooling method [RTC19]. This method is a versatile pooling mechanism used in CNNs for tasks such as image retrieval. It is based on the generalized-mean and incorporates learnable parameters, which can either be applied globally or individually for each output dimension. This flexibility allows GeM pooling to generalize both max pooling and average pooling as special cases, enabling it to adapt effectively to various tasks and improve performance over traditional pooling methods. For a spatial activation map X_k corresponding to the k-th feature, the GeM pooling function is defined as:

$$f_k^{(g)} = \left(\frac{1}{|X_k|} \sum_{x \in X_k} x_k^{p_k} \right)^{1/p_k},$$

where $|X_k|$ denotes the number of elements in X_k, and p_k is the pooling parameter. Special cases of GeM pooling include max pooling when $p_k \to \infty$ and average pooling when $p_k = 1$.

The differentiability of the GeM operation allows the pooling parameter p_k to be learned during back-propagation, enabling the network to optimize pooling behavior for specific tasks. This adaptability makes GeM pooling particularly effective for constructing compact

image descriptors, as it balances emphasizing the most significant activations and generalizing across the spatial feature map. GeM pooling has been shown to outperform traditional non-trainable pooling methods in tasks requiring robust image representations.

The FFNN serves as the classifier, outputting predictions across a set of unique concepts $\mathcal{C}$. The output layer consists of $|\mathcal{C}|$ neurons, each corresponding to a concept from the dataset. Each neuron employs a sigmoid activation function to produce a probability score between 0 and 1 for the associated concept. A concept is assigned to an image if its probability exceeds a predefined threshold t. This threshold, which is the same across all concepts, was determined via grid search on the validation set, optimizing the competition's primary evaluation metric, which in this case was the F_1-score. The model is trained by minimizing the binary cross-entropy loss, treating each concept as an independent binary classification target. The total loss is calculated by summing the individual binary losses across all concepts. Training is conducted using the Adam optimizer [KB17] with an initial learning rate $\eta = 10^{-3}$, and a learning rate decay strategy. Early stopping is employed, monitoring the validation loss with a patience of 3 epochs to prevent overfitting.

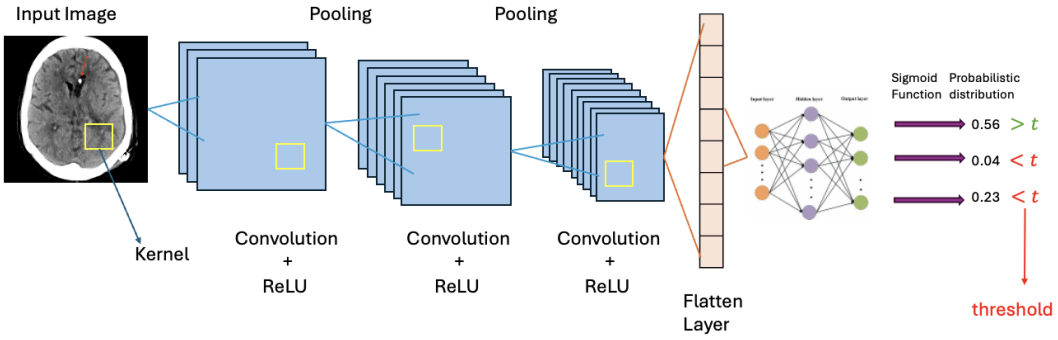


Fig. 3.1: The system uses a CNN for feature extraction and an FFNN for classification, with GeM pooling to generate image embeddings. Concepts are predicted using sigmoid probabilities, with a threshold t applied uniformly.

3.2 CNN encoder + k-NN retrieval

We experimented with various k-nearest neighbors (k-NN) approaches, including a weighted k-NN method [Sam+24].The method described here yielded the best results. We utilized image embeddings derived from the image encoder described in Section 3.1. After discarding the dense classification head, we extracted embeddings (also known as feature vectors) from the last GeM pooling layer [Cha+21]for all training images. These embeddings formed the foundation for the retrieval process of the k-NN algorithm. For each test image, we used the same encoder to generate its embedding and retrieved the k closest neighbors from the training set. The resemblance between images was determined by calculating the cosine similarity between their embeddings. From these retrieved neighbors, we selected

which concepts to assign to the test image as follows: We identified its k nearest neighbors from the training set and compiled the set of concepts associated with these neighbors. We then ranked these concepts based on their frequency among the retrieved neighbors. The concept with the highest frequency was always included in the predictions for the test image.

To determine which additional concepts should be included in the predictions, we applied two thresholds, t_1 and t_2, which were optimized using grid search on our validation set. Specifically, we computed the difference in frequency among the retrieved neighbors between the most frequent and the second most frequent concepts, normalized by the frequency of the most frequent concept. If this ratio was below t_1, we included the second concept in the prediction; otherwise we did not:

$$\frac{\mathbf{Fr}(\text{concept}_1) - \mathbf{Fr}(\text{concept}_2)}{\mathbf{Fr}(\text{concept}_1)} < t_1 \,. \tag{3.1}$$

Similarly, we evaluated whether to include the third most frequent concept in the prediction by comparing the frequencies of the first and third most frequent concepts. We calculated the difference in frequency between these two concepts, divided by the frequency of the first concept. If this ratio was below the threshold t_2, the third concept was included in the prediction as well:

$$\frac{\mathbf{Fr}(\text{concept}_1) - \mathbf{Fr}(\text{concept}_3)}{\mathbf{Fr}(\text{concept}_1)} < t_2 \,. \tag{3.2}$$

The same method was used to evaluate the difference between the frequencies of the first and fourth most frequent concepts. This difference was compared to the threshold t_2 to determine whether the fourth most frequent concept should be included in the prediction:

$$\frac{\mathbf{Fr}(\text{concept}_1) - \mathbf{Fr}(\text{concept}_4)}{\mathbf{Fr}(\text{concept}_1)} < t_2 \,. \tag{3.3}$$

We chose to predict up to four concepts because the average number of concepts in the training set was 3.08, as it is depicted in Figure 3.2. Our approach was to select concepts with frequencies similar to the most frequent concept, while excluding those with a significant drop in frequency compared to the previous ones. Initially, we considered using three thresholds to refine the predictions further. However, during the tuning process, we found that the third threshold consistently matched the second threshold's value, making it redundant. As a result, we simplified the approach to using only two thresholds, t_1

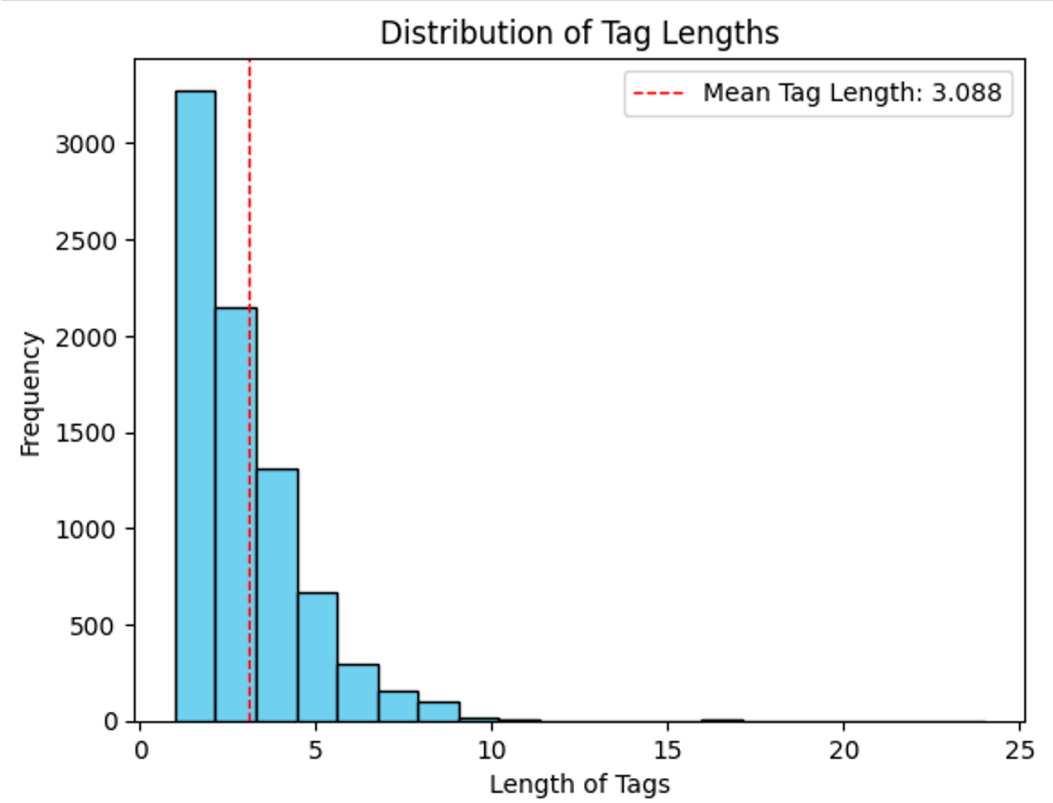


Fig. 3.2: Histogram illustrating the distribution of the number of tags (concepts) per image in the ImageCLEFmedical dataset [Rüc+24b]. Y-axis showing number of images and x-axis number of tags per image.

and t_2. We experimented with different values for t_1 and t_2 ranging from 0.3 to 0.9. The validation results showed that the optimal parameters were $t_1 = 0.58$ and $t_2 = 0.65$.

3.3 CNN-FFNN + k-NN ensemble

To address cases where the classification confidence of CNN-FFNN is low, we extend the CNN-FFNN model by incorporating a k-NN model. This hybrid architecture, termed CNN-FFNN + k-NN, leverages the feature extraction capabilities of the CNN and the classification power of the FFNN while refining predictions in uncertain regions using k-NN. Specifically, we introduce a two-threshold strategy with a "grey zone" for uncertain predictions.

3.3.1 Two-Threshold Decision Mechanism

In this enhanced model, concept assignment is determined by two thresholds, t_1 and t_2, where $t_2 \leq t_1$. This approach divides the prediction space into three distinct regions:

1. **Confident Predictions** ($p \geq t_1$): If a concept's predicted probability, as determined by the CNN-FFNN model, exceeds t_1, it is confidently assigned to the image based on the system described in Section 3.1.

2. **Grey Zone** ($t_2 \leq p < t_1$): For predictions that fall between t_1 and t_2, the model considers these concepts uncertain. This ambiguity is resolved by applying k-NN based on image embeddings.

3. **Unassigned Predictions** ($p < t_2$): If the predicted probability is below t_2, the concept is not assigned to the image.

3.3.2 Grey Zone Classification Using k-NN

When the predicted probability p for a concept lies within the grey zone (i.e., $t_2 \leq p < t_1$), the ensemble defers to k-NN to refine the decision. For each concept in the grey zone, k-NN retrieves the k nearest neighbors from the training set based on the cosine similarity between their embeddings and the embedding of the test image. The tags of these neighbors are then used to determine whether the concept should be assigned to the image. Specifically, if a concept within the grey zone is also included in the prediction set generated by the k-NN , it is included in the final set of predicted tags; otherwise it is not included. This process improves the model's ability to classify ambiguous cases by leveraging similar cases from the training data.

3.3.3 Threshold Tuning

To optimize the classification performance, we perform threshold tuning on the validation set. Specifically, we tune the two thresholds, t_1 and t_2, that determine the concept assignment process. The thresholds t_1 and t_2 are adjusted to maximize the F_1 score, which balances precision and recall across all concepts. The tuning process involves experimenting with all possible combinations of t_1 and t_2 within a predefined range (e.g., in $[0.1, 1]$ with a step size of 0.01) under the constraint $t_2 \leq t_1$.

Once all threshold combinations are evaluated, the pair that achieves the highest F_1 score on the validation set is selected as the optimal threshold configuration for the model.

3.4 Other Ensemble Systems

Subsequently, we implemented ensemble systems to improve the robustness and accuracy of our model predictions. Ensemble learning combines multiple models to produce a

stronger, with better generalization ability classifier, leveraging the strengths of different models to minimize individual model weaknesses. Our ensemble approach involved integrating predictions from multiple model instances, which were obtained through a combination of different model architectures as well as multiple variants of the same classifier generated using different checkpoints or hyper-parameter configurations, by aggregating their predicted concept sets. Two primary aggregation strategies were explored:

1. **Union of Predictions**: In this method, an image is assigned a concept if any of the models in the ensemble predicts it. This approach tends to maximize recall, as more concepts are included when combining predictions across models.

2. **Intersection of Predictions**: Here, a concept is only assigned to an image if all models in the ensemble predict it. This method prioritizes precision by ensuring that only confidently predicted concepts are retained.

By combining these strategies, we explored ensembles that utilized different combinations of Union and Intersection operations on the predicted concept sets. For example, we applied Union to merge predictions from two models, used Intersection to combine predictions from another pair, or combined both approaches—such as taking the Union of two models' predictions and then finding their Intersection with predictions from a third model. The various aggregation strategies and model combinations employed are detailed in Table 3.1, illustrating the methods used to balance precision and recall while leveraging the strengths of each model to meet the task-specific evaluation metrics. Our ensemble models consisted of combinations of different architectures, including CNN + FFNN and CNN + k-NN. The CNN + FFNN model used various CNN encoders (e.g., Densenet [Hua+17], EfficientNet [**efficientnet**; TL21]) as feature extractors, followed by an FFNN for classification.

Ensembles and Model Combinations
CNN+FFNN (DenseNet)
CNN+kNN
INTERSECTION(UNION(3xCNN+FFNN), CNN+kNN)
UNION(2xCNN+FFNN)
CNN+FFNN (EfficientNet)
UNION(CNN+FFNN (EfficientNet), CNN+kNN)
CNN+wkNN
UNION(CNN+wkNN, CNN+FFNN (EfficientNet))
UNION(CNN+wkNN, CNN+kNN)
UNION(CNN+wkNN, CNN+FFNN (DenseNet))

Tab. 3.1: Overview of all ensembles and model combinations explored in our experiments.

3.5 Discussion: Future Ensemble Strategies

While the ensemble strategies employed in this thesis demonstrated improvements in model performance, there remain several promising directions for future research that could further enhance robustness and accuracy. One such approach is "temporal averaging," where the weights of the best-performing checkpoints of a neural model are averaged to create a single, more stable model. This technique can stabilize performance by reducing the variability introduced by individual training iterations and has shown promise in reducing overfitting [LA16].

Another potential direction involves training models with more diverse hyper-parameter configurations, such as varying learning rates or dropout rates. Aggregating predictions from these models could lead to better generalization by leveraging the unique strengths of each configuration. Furthermore, ensemble methods could also involve weighted averaging, where predictions from different models are assigned weights proportional to their validation performance, creating a hybrid model that balances the strengths of each component.

Stacking is another promising avenue, where a meta-classifier is trained to learn how to best combine the predictions from different models. This meta-classifier can prioritize specific models based on their performance in particular cases, potentially improving performance in rare or challenging scenarios. Stacking has been widely studied and applied successfully in ensemble learning, as discussed in earlier works [Wol92].

In addition to these strategies, combining models with fundamentally different architectures could provide complementary benefits. For instance, integrating CNNs with retrieval-based methods or transformer-based architectures could enhance performance by exploiting the unique strengths of each approach. Such combinations could prove especially useful in tasks involving complex or imbalanced datasets.

These ensemble strategies have shown promise in previous ImageCLEF tasks. For example, Charalampakos et al. [Cha+22] utilized retrieval-based methods alongside deep learning classifiers to enhance medical image tagging. Similarly, Charalampakos et al. [Cha+21] explored ensembles of CNN encoders and feed-forward neural network classifiers to improve caption generation in medical datasets. Such approaches illustrate the potential of combining diverse methodologies to tackle complex problems effectively. Incorporating these advanced ensemble strategies in future research could lead to models with even greater robustness, adaptability, and accuracy, particularly for tasks involving complex or imbalanced datasets.

4 Data

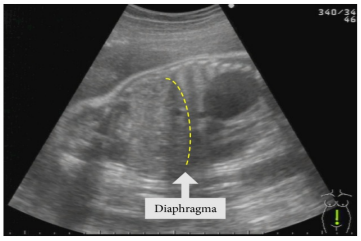
In the context of this thesis, we utilized the ImageCLEFmedical 2024 dataset [Rüc+24a], which offers a diverse collection of annotated medical images designed for tasks such as image classification, and retrieval. The following sections will provide a detailed overview of the ImageCLEFmedical 2024 dataset [Rüc+24b], along with an exploratory data analysis to better understand its characteristics, before applying the methods discussed in Chapter 3 to the data.

4.1 ImageCLEFmedical 2024

The ImageCLEFmedical Caption task [Rüc+24a] is part of the broader ImageCLEF competition [Ion+24]. It aims to advance the state of the art in image analysis and retrieval through challenging benchmarks that address real-world problems in the biomedical domain. In this year's ImageCLEFmedical Caption task, the dataset used is an enhanced and expanded version of the Radiology Objects in Context (ROCO) dataset [Pel+18], which is derived from biomedical articles within the PubMed Open Access (PMC OA) subset. This dataset, utilized for both sub-tasks, comprises 80,080 biomedical images, each accompanied by corresponding medical concepts represented as UMLS [Bod04] terms, along with detailed diagnostic captions. The two sub-tasks are Concept Detection, which involves identifying relevant medical concepts (tags) associated with each image, and Caption Prediction, which focuses on generating diagnostic captions describing the findings in the image. The original dataset provided by the organizers was divided into two subsets: a training set comprising 70,108 radiology images and a validation set containing 9,972 images. To better suit our evaluation needs, we merged these subsets and re-split the data into three distinct groups: training, validation, and a development (our test) subset. We implemented a $75\% - 10\% - 15\%$ split, ensuring that the distribution of concepts remained relatively balanced across all three subsets. As a result, our final split yielded 64,928 images for training, 7,179 images for validation, and 7,973 images designated as the development set (our own test set). Additionally, all our submissions were evaluated using the hidden official test set, derived from the Radiology Objects in Context Version 2 (ROCOv2) [Rüc+24b]. This updated and expanded version of the original ROCO dataset [Pel+18] includes 17,237 previously unseen images, providing a comprehensive and robust basis for testing.

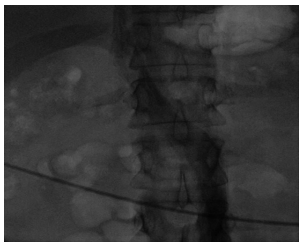
4.2 Concept Detection

Concept Detection in the ImageCLEFmedical Caption task is framed as a multi-class multi-label classification challenge, covering a wide array of 1,945 distinct biomedical concepts. These concepts are derived from the Unified Medical Language System (UMLS) [Bod04], which serves as a comprehensive repository of medical terminology. The primary objective of this sub-task is to accurately identify and assign the relevant medical concepts associated with each image. These concepts can range from specific medical conditions to anatomical structures, procedures, and imaging findings depicted within the images. Among the extensive set of available concepts, several are directly related to imaging modalities, including X-Ray Computed Tomography (CT), Ultrasonography, Magnetic Resonance Imaging (MRI), and Positron Emission Tomography (PET)/CT scans. Each concept within the dataset is represented by a unique Concept Unique Identifier (CUI) in accordance with the UMLS standard. In this task, each image may be associated with multiple concepts, meaning that an individual image can have several relevant tags reflecting different medical conditions presented in the dataset. Illustrative examples of images along with their corresponding ground truth concepts are presented in Figure 4.1.



CUI	UMLS Term
C0041618	Ultrasonography
C0025066	Mediastinum
C0003489	Aortic arch structure
CC BY [Yagi et al. (2018)]	

CUI	UMLS Term
C0040405	X-Ray CT
C0542331	Shunt Device
C0085590	catheter device
CC BY [Muacevic et al. (2021)]	



CUI	UMLS Term
C0002978	angiogram
C0015809	Femoral vein
C0018787	Heart
CC BY [Awan et al. (2020)]	

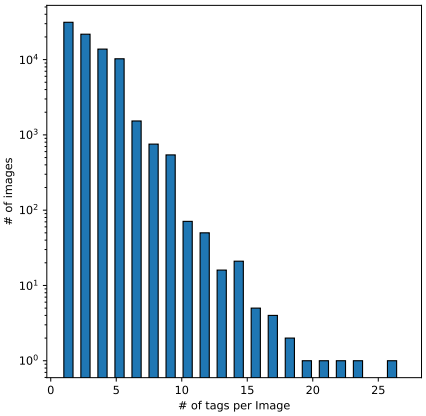
Fig. 4.1: Some example image-concept pairs from the ImageCLEFmedical2024 dataset. The concepts are presented in their UMLS term form.

The distribution of concepts in the dataset is highly imbalanced. While some concepts appear in over 25,000 images, others are linked to just a single image. Figure 4.2a illustrates this long-tail distribution for the entire dataset (development, validation, and training sets combined). The plot on the left shows the frequency of each concept (i.e., the number of images associated with each concept) in descending order, mapped against their respective class indices. Our thorough exploratory analysis of this year's dataset revealed that certain concepts are more frequently occurring (as seen in Table 4.1), and predominantly related to types of medical examinations, such as "X-Ray Computed Tomography" or "Plain X-ray". Most images in the ground truth are tagged with at least one of these broad concepts, in addition to more specialized ones. The number of concepts assigned to a single image

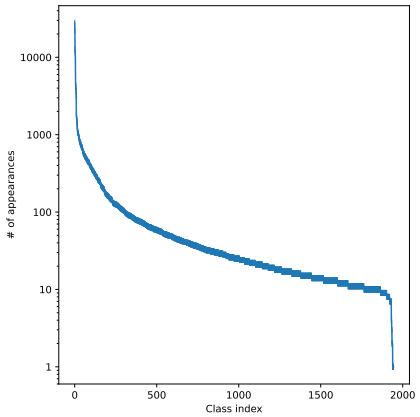
ranges from a minimum of 1 to a maximum of 27, with 8,567 images associated with just 1 concept, and only 1 image reaching the maximum. On average, each image is assigned 3.15 concepts. These observations are summarized in the histogram of Figure 4.2b.

Tab. 4.1: The ten most frequent concepts (CUIs) of the ImageCLEFmedical2024 dataset, along with their corresponding UMLS terms, and the number of images they are associated with.

Most Common Concepts			
Rank	**CUI**	**UMLS Term**	**Images**
1	C0040405	X-Ray Computed Tomography	27,852
2	C1306645	Plain x-ray	22,104
3	C0024485	Magnetic Resonance Imaging	12,733
4	C0041618	Ultrasonography	11,476
5	C0817096	Chest	10,323
6	C0002978	angiogram	4,808
7	C0000726	Abdomen	4,292
8	C0037303	Bone structure of cranium	4,130
9	C0030797	Pelvis	3,678
10	C0023216	Lower Extremity	3,254



(a) Number of tags per image.



(b) Long-tail distribution of concepts.

Fig. 4.2: (a) Histogram with 25 fixed-size bins (horizontal axis) depicting the number of gold concepts/tags per image. Note that 13 concepts do not have corresponding UMLS terms. (b) Visualization of the dataset's long-tail distribution. The y-axis shows the number of occurrences of each concept, and the x-axis the concept's class index.

5 Experiments and Results

In this chapter, we present the experiments conducted to evaluate the performance of the proposed methods for biomedical image tagging mentioned in Chapter 3.

The experimental setup includes several key phases, such as data pre-processing, model training, evaluation, and comparison with baseline models. We also investigate the impact of transfer learning techniques, as outlined in earlier sections, and compare the classification-based approach to the retrieval-based approach in terms of performance metrics such as PRECISION, RECALL and F_1 score.

5.1 Experimental setup

In this section, we describe in detail the setup used to run the experiments. We trained and evaluated the models using the ImageCLEFmedical 2024 dataset [Rüc+24a], which consists of medical images labeled with (multiple) biomedical concepts, as described in Chapter 4. The dataset was split into training, validation, and development sets with a 75%-10%-15% ratio.

5.1.1 Data Pre-processing

The pre-processing pipeline was designed to standardize the input data for robust model training. Each biomedical image was resized to a fixed dimension of 224×224 pixels, ensuring a consistent input size across the dataset. Pixel values were normalized using a preprocessor specific to the CNN backbone architecture employed (e.g., EfficientNetV2B0). This normalization adjusted the pixel intensity distribution to align with the expectations of the pre-trained model.

5.1.2 Transfer Learning

Transfer learning was applied by fine-tuning pre-trained models (EfficientNet [TL19; TL21], DenseNet [Hua+17]) on the medical images, leveraging their strong feature extraction capabilities to enhance performance on the specific task. The fine-tuning involved training all layers of the models, which were initialized with ImageNet weights, allowing the

encoders to adapt to the medical domain while preserving their pre-trained knowledge. Training was conducted for a maximum of 100 epochs, with early stopping applied to halt training if the validation loss did not improve for 3 consecutive epochs. A learning rate scheduler reduced the learning rate by a factor of 0.1 after one epoch of plateaued validation loss, ensuring effective convergence. The models were optimized using the Adam optimizer with an initial learning rate of 1×10^{-4} and a batch size of 16. Dropout layers and global pooling strategies, such as Generalized Mean Pooling (GeM), were employed to mitigate overfitting and aggregate features effectively. Performance monitoring relied on validation loss during training, while post-training evaluation included metrics such as F1 score to ensure a balanced trade-off between precision and recall.

5.1.3 Model Training

The models were trained [KSH12] using both CNN-based and retrieval-based approaches, as discussed in Section 3. For the CNN-based method, we employed CNN encoder networks combined with fully connected layers, optimized for multi-label classification. For the retrieval-based method, k-NN retrieval was performed using image embeddings extracted from CNN models.

5.2 Perfomance Evaluation Metrics

To evaluate the performance of our systems, we used the F_1 score as the primary metric, which balances both precision and recall in multi-label classification tasks. This approach is particularly useful for datasets with imbalanced label distributions, such as ImageCLEFmedical. The F_1 score was the official evaluation metric used in the competition, and all results were derived by calculating the F_1 score for each test image and then averaging over the test set. Specifically, the binary multi-hot predicted tags vector (y_{pred}) was compared to the ground truth vector (y_{true}) for each image, and $\hat{f}_1$[1] denotes the individual F_1 scores that were summed and averaged to obtain the final result. The F_1 score calculation follows:

$$F_1 = \frac{1}{|T|} \sum_{t \in T} \hat{f}_1(y_{pred}, y_{true}) \tag{5.1}$$

where (y_{pred}) and (y_{true}) denote the predicted and ground truth concepts for image t, and T is the set of test images. In addition to the primary evaluation metric, a secondary F1-score was computed, which focused solely on manually selected concepts, including categories such as anatomy, topography, and imaging modality. For this metric, predictions

[1] `https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html`

and ground truth concept lists were filtered to include only the selected concepts. Images with no relevant ground truth concepts after filtering were excluded from the evaluation. The F_1 score for each image was calculated by comparing binary vectors representing the presence or absence of each concept, and the overall score was obtained by averaging these per-image F_1 scores.

5.3 Results

For the Concept Detection sub-task, we identified our ten best-performing models by evaluating them on the held-out development (test) set, and these models were then submitted for evaluation on the official test set. The models consisted of various CNN-based architectures paired with different techniques for classification, including CNN + FFNN, CNN + k-NN, and CNN + weighted k-NN. Additionally, we employed ensemble systems, combining predictions from multiple instances of these models by either taking the union or intersection of their predicted concepts.

For our first system (CNN + FFNN) described in Section 3, we experimented with several CNN architectures as the classifier's backbone. Specifically, we experimented with EfficientNet [TL19; TL21] and DenseNet [Hua+17] to train the networks. These same CNN encoders were also incorporated into our k-NN models to further extend our experiments.

Testing on the held-out development set revealed that models utilizing the EfficientNet [TL19] image encoder achieved a marginally higher F_1-score compared to other architectures. Despite trying ensemble approaches, which combined predictions from multiple models, the results did not show significant improvement over the individual models, with only minor differences observed between the development and test sets [Rüc+24b].

Tab. 5.1: Summary of our submissions to the ImageCLEFmedical2024 Concept Detection sub-task. The table presents the primary F_1-scores of our systems on both our held-out development set and the official test set. The rankings of our systems among all 38 submissions from the 9 participating teams are included and are based on the primary F_1-score on the official test set. For the secondary F_1-score, only results on the official test set are included.

Individual Concept Detection Experiments					
Run ID	Method	Primary F1		Secondary F1	Rank
		Dev	Test		
619	CNN+FFNN (DenseNet)	0.6007	0.6240	0.9339	12
624	CNN+kNN	0.6007	0.6274	0.9375	8
640	INTERSECTION(UNION(3xCNN+FFNN),624)	0.6022	0.6272	**0.9415**	10
642	UNION(2xCNN+FFNN)	**0.6047**	0.6304	0.9332	7
644	CNN+FFNN (EfficientNet)	0.6042	**0.6319**	0.9392	**4**
648	UNION(644,624)	0.6045	0.6308	0.9321	6
651	CNN+wkNN	0.5961	0.6135	0.9238	17
654	UNION(651,644)	0.6008	0.6207	0.9243	13
655	UNION(651,624)	0.5970	0.6155	0.9233	16
656	UNION(651,619)	0.5981	0.6162	0.9217	15

5.4 Discussion

The results from our experiments reveal several key insights about the performance of different models for the Concept Detection task. Overall, our systems performed exceptionally well, achieving 2nd place among the 9 participating teams in the Concept Detection sub-task and securing 4th place overall across all submissions from these teams.

First, models utilizing the EfficientNet [TL19; TL21] image encoder consistently outperformed those with other CNN backbones, including DenseNet [Hua+17]. In contrast,our use of data augmentation techniques, which are typically employed to enhance generalization and prevent overfitting, did not yield significant improvements. This may suggest that the medical images in our dataset already contained enough variability, or that the complex nature of biomedical images limited the effectiveness of standard augmentation strategies.

Although our k-NN retrieval-based models performed well, especially in cases involving rare or less frequent concepts, they did not consistently surpass the CNN + FFNN models in the overall F_1-score. However, the performance of the k-NN models was closely comparable to the CNN + FFNN models, indicating that retrieval-based methods can be competitive for multi-label image classification. In fact, our k-NN models showed improvement over last year's approaches, largely due to the way we fine-tuned the number of tags included in the predictions, suggesting that retrieval-based techniques have evolved and now offer a viable alternative. There is still room for further exploration, particularly in optimizing similarity measures or incorporating more advanced embedding techniques to enhance the retrieval process and potentially outperform CNN-based classification approaches in future iterations. The ensemble methods, which combined predictions from multiple models, did not result in significant performance gains. This suggests that while individual models were already making strong predictions, the current ensemble approach may not have fully leveraged the potential benefits of combining models. Exploring alternative ensembling methods, such as combining models with more diverse CNN architectures, could offer better synergy between models and potentially improve robustness and accuracy.

6 Conclusions and Future Work

6.1 Conclusions

This thesis explored several deep learning techniques to address the challenges of biomedical multi-label image classification. The problem is critical in modern healthcare, where automated systems are increasingly necessary to handle the growing volume of medical images. By utilizing a combination of classification-based and retrieval-based methods, we aimed to enhance the accuracy and reliability of medical image tagging.

We implemented a variety of model architectures, including a CNN-FFNN probabilistic classifier and retrieval-based k-NN methods, and combined these in ensemble systems to further improve performance. Through rigorous experimentation on the ImageCLEFmedical 2024 dataset [Rüc+24a], we demonstrated that:

- **EfficientNet-based models** outperformed other architectures in terms of F_1-score, demonstrating the effectiveness of scaling methods that balance network depth, width, and resolution.

- **Retrieval-based methods (k-NN)**, while competitive, k-NN-based models did not consistently surpass CNN-based models but demonstrated potential, leaving room for further improvement in the future.

- **Ensemble models**, which aggregated predictions from multiple systems, did not yield significant performance improvements, indicating that alternative approaches should be explored. Strategies like temporal averaging, weighted averaging, or stacking, as well as integrating models with diverse architectures, could enhance robustness and accuracy in future work.

Finally, we achieved competitive results in the ImageCLEFmedical 2024 competition, ranking 2^{nd} in the Concept Detection task, validating the effectiveness of our proposed systems.

6.2 Future Work

In the future, there is potential to further explore retrieval-based methods and integrate more advanced architectures. While retrieval methods like k-NN showed promise, improvements can still be made by refining similarity metrics or combining them with deep learning models, as discussed in Chapter 3. Future work should also investigate the use of data augmentation techniques to address class imbalances and improve model generalization, especially for under-represented concepts.

Bibliography

[AMT22] Ioannis Athanasiadis, Georgios Moschovis, and Alexander Tuoma. "Weakly-Supervised Semantic Segmentation via Transformer Explainability". In: *Proceedings of the ML Reproducibility Challenge 2021 (Fall Edition)*. 2022.

[BB22] Samiya Majid Baba and Indu Bala. "Detection of Diabetic Retinopathy with Retinal Images using CNN". In: *26th International Conference on Intelligent Computing and Control Systems (ICICCS)*. 2022, pp. 1074–1080.

[Bod04] Olivier Bodenreider. "The Unified Medical Language System (UMLS): Integrating Biomedical Terminology". In: *Nucleic acids research* 32 (Feb. 2004), pp. D267–70.

[CH67] Thomas Cover and Peter Hart. "Nearest neighbor pattern classification". In: *IEEE Transactions on Information Theory* 13.1 (1967), pp. 21–27.

[Cha+21] Foivos Charalampakos, Vasilis Karatzas, Vasiliki Kougia, John Pavlopoulos, and Ion Androutsopoulos. "AUEB NLP Group at ImageCLEFmed Caption Tasks 2021". In: *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21-24*. Vol. 2936. CEUR Workshop Proceedings. 2021, pp. 1184–1200.

[Cha+22] Foivos Charalampakos, George Zachariadis, John Pavlopoulos, et al. "AUEB NLP Group at ImageCLEFmedical Caption 2022". In: *CLEF2022 Working Notes*. CEUR Workshop Proceedings. Bologna, Italy: CEUR-WS.or, 2022, pp. 1355–1373.

[Che+19] Chen Chen, Xiaopeng Liu, Meng Ding, Junfeng Zheng, and Jiangyun Li. "3D Dilated Multi-Fiber Network for Real-time Brain Tumor Segmentation in MRI". In: *arXiv preprint arXiv:1904.03355* (2019).

[DBK+21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *Proceedings of the 9th International Conference on Learning Representations (ICLR)*. Vienna, Austria, 2021.

[Den+09] Jia Deng, Wei Dong, Richard Socher, et al. "ImageNet: A large-scale hierarchical image database". In: *IEEE Conference on Computer Vision and Pattern Recognition*. 2009, pp. 248–255.

[Dev+19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186.

[DN21] Prafulla Dhariwal and Alex Nichol. "Diffusion Models Beat GANs on Image Synthesis". In: *arXiv preprint arXiv:2105.05233* (2021).

[Dos+21] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations*. 2021.

[DT05] Navneet Dalal and Bill Triggs. "Histograms of oriented gradients for human detection". In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 1. IEEE. 2005, pp. 886–893.

[Eic+17] Carsten Eickhoff, Iris Schwall, Alba García Seco de Herrera, and Henning Müller. "Overview of ImageCLEFcaption 2017 - the Image Caption Prediction and Concept Extraction Tasks to Understand Biomedical Images". In: *CLEF2017 Working Notes*. Dublin, Ireland: CEUR Workshop Proceedings, 2017.

[He+16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 2016, pp. 770–778.

[Hua+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, pp. 2261–2269.

[Ift+24] Tanzina Taher Ifty, Saleh Ahmed Shafin, Shoeb Mohammad Shahriar, and Tashfia Towhid. "Explainable Lung Disease Classification from Chest X-Ray Images Utilizing Deep Learning and XAI". In: *arXiv preprint arXiv:2404.11428* (2024).

[Ion+24] Bogdan Ionescu, Henning Müller, Ana-Maria Drăgulinescu, et al. "Overview of ImageCLEF 2024: Multimedia Retrieval in Medical Applications". In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Proceedings of the 15th International Conference of the CLEF Association (CLEF 2024). Grenoble, France: Springer Lecture Notes in Computer Science LNCS, Sept. 2024.

[Irv+19] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, et al. *CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison*. 2019. arXiv: 1901.07031 [cs.CV].

[JS22] Qiran Jia and Hai Shu. "BiTr-Unet: a CNN-Transformer Combined Network for MRI Brain Tumor Segmentation". In: *MICCAI BrainLes 2021: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Vol. 12963. Springer, 2022, pp. 3–14.

[Kal+23] Panagiotis Kaliosis, Georgios Moschovis, Foivos Charalampakos, John Pavlopoulos, and Ion Androutsopoulos. "AUEB NLP Group at ImageCLEFmedical Caption 2023". In: *Conference and Labs of the Evaluation Forum (CLEF)*. Thessaloniki, Greece, 2023.

[Kal+24] Panagiotis Kaliosis, John Pavlopoulos, Foivos Charalampakos, Georgios Moschovis, and Ion Androutsopoulos. "A Data-Driven Guided Decoding Mechanism for Diagnostic Captioning". In: *Findings of the Association for Computational Linguistics ACL 2024*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, Aug. 2024, pp. 7450–7466.

[Kar+20] Vasilis Karatzas, Vasiliki Kougia, John Pavlopoulos, and Ion Androutsopoulos. "AUEB NLP Group at ImageCLEFmed Caption 2020". In: *Conference and Labs of the Evaluation Forum (CLEF)*. 2020.

[Kas+24] Idan Kassis, Dror Lederman, Gal Ben-Arie, et al. "Detection of breast cancer in digital breast tomosynthesis with vision transformers". In: *Scientific Reports* 14 (2024), p. 22149.

[KB17] Diederik P. Kingma and Jimmy Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG].

[Kha+22] Firas Khader, Gustav Mueller-Franzes, Soroosh Tayebi Arasteh, et al. "Medical Diffusion: Denoising Diffusion Probabilistic Models for 3D Medical Image Generation". In: *arXiv preprint arXiv:2211.03364* (2022).

[KPA19] Vasiliki Kougia, John Pavlopoulos, and Ion Androutsopoulos. "AUEB NLP Group at ImageCLEFmed Caption 2019". In: *CLEF2019 Working Notes*. CEUR Workshop Proceedings. Lugano, Switzerland: CEUR-WS.org, 2019.

[KPA20] Vasiliki Kougia, John Pavlopoulos, and Ion Androutsopoulos. "Medical Image Tagging by Deep Learning and Retrieval". In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 11th International Conference of the CLEF Association (CLEF 2020)*. Springer International Publishing, 2020, pp. 154–166.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*. Vol. 25. 2012.

[LA16] Samuli Laine and Timo Aila. "Temporal Ensembling for Semi-Supervised Learning". In: *arXiv preprint arXiv:1610.02242* (2016).

[LD15] Shuying Liu and Weihong Deng. "Very deep convolutional neural network based image classification using small training sample size". In: *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*. 2015, pp. 730–734.

[LeC+89] Yann LeCun, Bernhard Boser, John S. Denker, et al. "Backpropagation applied to handwritten zip code recognition". In: *Neural Computation* 1.4 (1989), pp. 541–551.

[LeC+98] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324.

[Lee+20] Jinsoo Lee, Jiwon Kang, Soohyun Park, Doyoung Jang, and Jaewook Lee. "A Multi-Class Classification Model for Technology Evaluation". In: *Sustainability* 12.15 (2020), pp. 1–16.

[Lit+17] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, et al. "A survey on deep learning in medical image analysis". In: *Medical Image Analysis* 42 (2017), pp. 60–88.

[Low04] David G. Lowe. "Distinctive image features from scale-invariant keypoints". In: *International Journal of Computer Vision* 60.2 (2004), pp. 91–110.

[MF22] Georgios Moschovis and Erik Fransén. "NeuralDynamicsLab at ImageCLEF Medical 2022". In: *CLEF2022 Working Notes*. CEUR Workshop Proceedings. Bologna, Italy: CEUR-WS.org, Sept. 2022.

[MNA16] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation". In: *arXiv preprint arXiv:1606.04797* (2016).

[Mos22] Georgios Moschovis. "Medical image captioning based on Deep Architectures". Last accessed: 2023-07-07. MA thesis. Stockholm, Sweden: KTH Royal Institute of Technology, Dec. 2022.

[Naj22] Reabal Najjar. "Redefining Radiology: A Review of Artificial Intelligence Integration in Medical Imaging". In: *Journal of Radiology* (2022). Ed. by Michał Strzelecki, Adam Piorkowski, and Rafał Obuchowicz.

[PBG+21] Otto Pelka, Asma Ben Abacha, Alba García Seco de Herrera, et al. "Overview of the ImageCLEFmed 2021 Concept & Caption Prediction Task". In: *CLEF2021 Working Notes*. Bucharest, Romania: CEUR Workshop Proceedings, 2021.

[Pel+18] Oliver Pelka, Stefan Koitka, Jonas Rückert, Frederik Nensa, and Christoph Friedrich. "Radiology Objects in COntext (ROCO): A Multimodal Image Dataset". In: *7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018*. Vol. 11043. Lecture Notes in Computer Science. Granada, Spain: Springer, 2018, pp. 180–189.

[Pel+19] Otto Pelka, Christoph M. Friedrich, Alba García Seco de Herrera, and Henning Müller. "Overview of the ImageCLEFmed 2019 Concept Prediction Task". In: *CLEF2019 Working Notes*. Vol. 2380. Lugano, Switzerland: CEUR Workshop Proceedings, 2019.

[Pel+20] Otto Pelka, Christoph M. Friedrich, Alba García Seco de Herrera, and Henning Müller. "Overview of the ImageCLEFmed 2020 Concept Prediction Task: Medical Image Understanding". In: *CLEF2020 Working Notes*. Vol. 1166. Thessaloniki, Greece: CEUR Workshop Proceedings, 2020.

[Rah+20] Tawsifur Rahman, Muhammad E. H. Chowdhury, Amith Khandakar, et al. "Transfer Learning with Deep Convolutional Neural Network (CNN) for Pneumonia Detection using Chest X-ray". In: *Applied Sciences* 10.9 (2020), p. 3233. arXiv: arXiv:2004.06578 [eess.IV].

[Raj+17] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, et al. "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning". In: *arXiv preprint arXiv:1711.05225* (2017).

[RBG+22] Jonas Rückert, Asma Ben Abacha, Alba García Seco de Herrera, et al. "Overview of ImageCLEFmedical 2022 - Caption Prediction and Concept Detection". In: *CLEF2022 Working Notes*. Bologna, Italy: CEUR Workshop Proceedings, 2022.

[RHW86] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors". In: *Nature* 323 (1986), pp. 533–536.

[Ros57] Frank Rosenblatt. *The Perceptron: A Perceiving and Recognizing Automaton*. Tech. rep. Cornell Aeronautical Laboratory, 1957.

[Ros62] Frank Rosenblatt. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Washington, DC: Spartan Books, 1962.

[RTC19] Filip Radenović, Giorgos Tolias, and Ondřej Chum. "Fine-Tuning CNN Image Retrieval with No Human Annotation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.7 (2019), pp. 1655–1668.

[Rüc+24a] Johannes Rückert, Asma Ben Abacha, Alba G. Seco de Herrera, et al. "Overview of ImageCLEFmedical 2024 – Caption Prediction and Concept Detection". In: *CLEF2024 Working Notes*. CEUR Workshop Proceedings. Grenoble, France: CEUR-WS.org, Sept. 2024.

[Rüc+24b] Jonas Rückert, Lars Bloch, Robert Brüngel, et al. "ROCOv2: Radiology Objects in COntext Version 2, an Updated Multimodal Image Dataset". In: *Scientific Data* (2024). URL: https://arxiv.org/abs/2405.10004v1.

[Rus+15] Olga Russakovsky, Jia Deng, Hao Su, et al. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision* 115.3 (Dec. 2015), pp. 211–252.

[Sam+24] Marina Samprovalaki, Anna Chatzipapadopoulou, Georgios Moschovis, et al. "AUEB NLP Group at ImageCLEFmedical Caption 2024". In: *Notebook for the AUEB NLP Group at ImageCLEFmedical Caption 2024, Conference and Labs of the Evaluation Forum (CLEF)*. Athens, Greece, 2024.

[Sel+20] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization". In: *International Journal of Computer Vision (IJCV)* 128 (2020), pp. 336–359. arXiv: arXiv:1610.02391 [cs.CV].

[Shi+18] Hoo-Chang Shin, Neil A. Tenenholtz, Jameson K. Rogers, et al. “Medical Image Synthesis for Data Augmentation and Anonymization using Generative Adversarial Networks”. In: *arXiv preprint arXiv:1807.10225* (2018).

[Taj+20] Nima Tajbakhsh, Leo Jeyaseelan, Quanzeng Li, et al. “Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation”. In: *Medical Image Analysis* 63 (2020), p. 101693.

[TL19] Mingxing Tan and Quoc Le. “EfficientNet: Rethinking model scaling for convolutional neural networks”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. Vol. 97. Long Beach, California, USA: PMLR, 2019, pp. 6105–6114.

[TL21] Mingxing Tan and Quoc V. Le. “EfficientNetV2: Smaller Models and Faster Training”. In: *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*. Vol. 139. Proceedings of Machine Learning Research. 2021, pp. 10096–10106.

[Vas+17a] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. Von Luxburg, S. Bengio, et al. Vol. 30. Curran Associates, Inc., 2017.

[Vas+17b] Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. “Attention is all you need”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 30. Curran Associates, Inc., 2017, pp. 5998–6008.

[VS17] Leonidas Valavanis and Spyridon Stathopoulos. “IPL at ImageCLEFmed 2017 Concept Prediction Task”. In: *CLEF2017 Working Notes*. Dublin, Ireland: CEUR Workshop Proceedings, 2017.

[Wol92] David H. Wolpert. “Stacked generalization”. In: *Neural Networks* 5.2 (1992), pp. 241–259.

[Zhu+20] Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, et al. “A Comprehensive Survey on Transfer Learning”. In: *arXiv preprint arXiv:1911.02685* (2020).

[ZLZ20] Xiang Zhou, Min Li, and Xiangwei Zeng. “Hierarchical Classification for Medical Imaging Using Few-Shot Learning Techniques”. In: *IEEE Access* 8 (2020), pp. 177823–177835.

[ZWL22] Hao Zheng, Guohui Wang, and Xuchen Li. “Swin-MLP: a strawberry appearance quality identification method by Swin Transformer and multi-layer perceptron”. In: *Journal of Real-Time Image Processing* 16.4 (2022), pp. 1–12.

List of Acronyms

CNN Convolution Neural Network

FFNN Feed Forward Neural Network

k-NN k-Nearest Neighbors

MRI Magnetic Resonance Imaging

CT Computed Tomography

PET Positron Emission Tomography

ViTs Vision Transformers

MLPs Multi-Layer Perceptrons

UMLS Unified Medical Language System

NLP Natural Language Processing

TL Transfer Learning

LM Language Model

SIFT Scale-Invariant Feature Transform

HOG Histogram of Oriented Gradients

DMMCS Distance From Median Maximum Cosine Similarity

SVMs Support Vector Machines

BoC Bag of Colours

BoVW Bag of Visual Words

MCML Multi-class, Multi-label classification

SoTA state-of-the-art

List of Figures

List of Tables