UNIVERSITÀ
DI PAVIA

Bachelor Thesis

Fusing Vision and Language Models for Radiographic Diagnostics: A Proof of Concept

Bahenda Yvon Dylan Ntegano
Artificial Intelligence

Academic Year 2024/2025

Bachelor in Artificial Intelligence
Pavia 2025

Bachelor in Artificial Intelligence

Bahenda Yvon Dylan Ntegano

Fusing Vision and Language Models for Radiographic Diagnostics: A Proof of Concept

Supervisor: **Prof. Federico Cabitza, University of Milano-Bicocca**

“A picture is worth a thousand words, but a few words can change its story.”

Sebastyne Young

To my family and friends, for their endless support; To Prof. Cabitza, for his guidance; And to Dr. Gallazzi, for his valued contribution.

Yours Truly

Abstract

Clinical decisions are made based on information from multiple modalities, like images and text, yet most AI systems process these modalities alone. This thesis investigates multimodal fusion, combining information from images with that from texts, to see if it leads to better performance in diagnosis, in medical imaging tasks.

The thesis project examines this hypothesis through a scientific investigation using two clinical cases. The first case is a multi-label chest X-ray classification with 14 abnormality labels, while the second case is a binary fracture detection task using appendicular skeletal X-rays. To this end, distilled Transformer models were used, a Vision Transformer (`DeiT`) for images and a BERT-based model (`TinyBERT`) for text. Various multimodal approaches were tested, including `CLIP` and several fusion mechanisms such as concatenation, projection with gating, and cross-attention. Due to limited computational resources, all models were frozen during training.

The results show that multimodal approaches improve diagnostic performance over single-modality baselines. In addition, language-only models generally outperform vision-only models. Multimodal approaches show better generalisation capabilities, especially `CLIP` and the fusion technique via projection with gating. It can be seen that projection with gating is a good compromise between simplicity and performance across the two clinical cases.

These findings validate the central hypothesis of this thesis that the combination of visual and textual information leads to more effective diagnostic performance in medical imaging tasks. Although the work is considered as a scientific investigation, it is a stepping stone towards the future development of multimodal clinical decision support systems.

Keywords

Multimodal AI, Vision Transformer, Language Model, Fusion, Chest X-rays, Appendicular Skeletal X-rays

Contents

Chapter 1
Introduction

Artificial intelligence is rapidly becoming a vital part in health care, especially in the field of medical imaging. With radiographic data becoming increasingly complex and increasing in volume, as well as the shortage of expert radiologists, there is a need for automated clinical decision support systems that can provide fast and accurate diagnoses. Deep learning methods, particularly Transformer-based models, have become increasingly common in recent years due to their robust performance. Although they were originally developed for applications in natural language processing, Transformers have been extended to other application areas and have been effective in computer vision in particular, which has led to their adoption in medical imaging research.

In radiology, tasks such as diagnosing conditions from chest or skeletal X-rays usually require analysing both image data and clinical context (in the form of radiology reports). To reach a conclusion (diagnosis), radiologists combine the information from the image with information about the patient's history, symptoms, indication, and findings. This link between modalities is key for high-level reasoning, which makes single-modality AI decision support systems lacking, since they operate on the input modalities in isolation. To overcome this limitation, multimodal models come into play. Multimodal models learn from visual and textual input at the same time, to better reflect the high-level reasoning of clinicians in practice.

1.1 The Role of AI in Medical Imaging

Integrating artificial intelligence into radiology addresses urgent issues such as variability in diagnosis, increase in workload, and the demand for more experts in the field. Among the available AI approaches, Transformer-based architectures in particular are promising since they can learn long-range dependencies and they can generalise better across different modalities and tasks. They have been successful with both natural language processing and computer vision, making them suitable for medical imaging, where understanding complicated patterns in both visual and textual information is crucial.

For this thesis, the focus is on Transformers and this was mainly influenced by data availability along with the recent trend in literature, as well as architectural superiority. Transformers perform better in many tasks and are robust in handling different input types compared to standard convolutional or recurrent neural networks. Above all, their transfer learning ability makes them especially useful in low-resource clinical settings where labelled data may be scarce.

Most importantly, this thesis adopts the perspective that AI should serve in a strictly assistive role; not as a tool to replace radiologists, but as a tool to help them provide better patient care. This is important in the context of medical imaging, in particular because diagnoses are often not based on image data alone. Clinical reports usually have essential and complementary information to images like patient history, indication, findings, and impressions. Radiologists integrate this multimodal

information to make diagnostic decisions, something that current unimodal AI systems are unable to replicate.

Diagnostic decisions in real-world clinical settings depend on a variety of factors, including physical exams, laboratory results, previous imaging (or medical history in general), and physician reports. Although the focus of this thesis is on the fusion of radiographic images and their corresponding radiology reports, ultimately the goal is to build AI clinical decision support systems that are more closely aligned with clinical processes. Fusing textual information with visual information significantly improves performance. Furthermore, applying explainability techniques to models can provide additional insights into the decision-making process of the model.

Multimodal AI systems, therefore, have the ability to improve the precision and trustworthiness of clinical decision support systems in medical diagnostics. However, there are still significant limitations, such as the lack of structured datasets and the high computational costs of deploying fusion-based models in clinical settings. In spite of these limitations, multimodal Transformers bring us a step closer to building AI systems that mimic the clinical reasoning process; by integrating information, not isolating it.

1.2 Problem Formulation

This thesis examines automatic disease classification using medical imaging data, such as chest and appendicular skeletal radiographs, together with their corresponding clinical reports. The classification task is a supervised approach; the goal is to predict one or more abnormal findings or lack thereof, given an X-ray image and the associated radiology report. The decision to adopt a classification task was mainly based on the prevalence and clinical usefulness of classification in current clinical decision support systems.

In order to evaluate the proposed methods, two clinical cases were evaluated. The first case was chest radiographs, which is a very common imaging modality and, as a result, has high data availability; the X-ray data was obtained from the publicly accessible NLMCXR dataset. The second case was about appendicular skeletal X-rays that were provided by a doctor at the Gaetano Pini Orthopedic Institute in Milan, offering a real-world setting and a contrasting clinical domain. This setup, using two cases, allowed the evaluation of the adaptability and generalisability across domains of the models that were proposed. In the case of chest X-rays, the task was multi-label, where each image and text pair could potentially have multiple abnormalities. Regarding the skeletal case, the task was a binary classification that predicted the presence or absence of a fracture or dislocation.

The main goal of this thesis is to evaluate whether combining visual and textual information leads to improved diagnostic performance. The specific objectives are as follows.

- Implementation of unimodal baselines using a Vision Transformer (`DeiT`) and a BERT-based Transformer (`TinyBERT`);
- Implementation and comparison of various multimodal fusion techniques, such as concatenation, projection with gating, multi-head cross-attention, and `CLIP`;
- Evaluation of the adaptability and generalisability of each model across both chest and skeletal tasks.

Although multimodal fusion was hypothesised to perform better, this thesis approaches the question as an open research question. Based on literature and clinical intuition, the expectation was that the fusion of visual and textual input would provide a better understanding and, therefore, lead to an increased diagnostic accuracy.

The central hypothesis of this thesis is that the combination of visual and textual information leads to more effective diagnostic performance in medical imaging tasks compared to unimodal approaches. By testing this hypothesis across two different clinical domains using frozen, Transformer-based

architectures, this thesis takes a step towards building adaptable and clinically relevant decision support systems.

1.3 Outline

The rest of this thesis is organised as follows.

Chapter 2 Related work, including background and SOTA;

Chapter 3 Methodology, including datasets and models used;

Chapter 4 Experiments performed across two clinical use cases;

Chapter 5 Results and Discussion;

Chapter 6 Conclusions.

Chapter 2
Related Work

2.1 Background

2.1.1 Medical Imaging

X-rays were discovered on the 8th of November in 1895 by Wilhelm Conrad Röntgen as a novel form of electromagnetic radiation. Röntgen referred to them as "X" rays to show that they were of an unknown nature. Within a year of his discovery they were already being applied to the clinical domain, with applications ranging from diagnosis to therapy.

Today, X-ray imaging is still one of the most used tools for diagnosis in medicine. It can be applied to various regions of the body, including the thorax (chest X-rays), the appendicular skeleton (arms and legs), abdomen, and spine.

Chest X-ray (CXR) is the most common radiological imaging worldwide used to diagnose, detect, and monitor thoracic diseases. It provides rapid, low-cost imaging of the heart, lungs, and bones. CXR, or chest radiography, is based on the differential absorption of X-ray photons by tissues of varying densities to produce 2D images (projections) of internal structures. Standard projections are posteroanterior (PA) and lateral projections, but when the patient is too sick to stand, anteroposterior (AP) projections are used instead. Most modern systems use digital radiography because of its greater contrast resolution, immediate image review, and improved storage capabilities compared to traditional screen-film systems.

Similarly, appendicular skeletal radiographs are widely used to diagnose fractures, dislocations, joint abnormalities, and bone lesions. They are typically acquired in standardised projections, like PA and lateral, depending on the region and purpose. They also benefit from modern digital imaging systems.

Modern radiography interpretation workflows incorporate digital images through the use of Picture Archiving and Communication Systems (PACS) and Radiology Information Systems (RIS) to streamline image acquisition and display. PACS handles medical image storage and retrieval according to the DICOM standard, while RIS handles report generation and links images to patient records, usually using Electronic Health Records (EHRs) for accessibility and traceability.

Radiology reports translate X-ray findings into clinical suggestions and follow standard sections: Indication, Technique, Findings, and Impression, and some systems also include a comparison and clinical history. Standardisation of clinical reports ensures clarity and supports decision making. Despite the advantage of structure reporting, free text reports are still dominant, causing difficulties for natural language processing (NLP) and big data mining. Free-text reports are still the norm since they allow flexibility and efficiency on the part of radiologists. Recent proposals based on transformer-based models attempt to convert free text reports to structured templates, but broader adoption is limited by workflow integration and usability concerns.

2.1.2 Deep Learning Fundamentals

A variety of deep learning architectures have significantly advanced the fields of computer vision and natural language processing over the past several years. These architectures range from pure attention-based Transformers to lightweight distilled models and multimodal contrastive learners. In this subsection, I focus on some of the pillars of that transformation that are relevant to this thesis. Vision Transformers (with a focus on `DeiT`), BERT-based models (in particular `TinyBERT`), the attention mechanism which is at the core of these architectures, and the multimodal framework `CLIP`. These models and concepts represent the fundamental components for understanding and designing deep learning systems that are capable of interpreting medical images and associated clinical textual reports.

Vision Transformers and DeiT

Traditional Convolutional Neural Networks (CNNs) have dominated image classification for a long time, and they do a great job in learning local spatial hierarchical features. Recently, research has shown that self-attention mechanisms that were originally designed for natural language processing perform equally on vision tasks. Vision Transformers (ViTs), introduced by Dosovitskiy et al. (2020) [12], adapt the Transformer architecture to images by segmenting an image into non-overlapping patches, flattening and processing them as sequences forming embeddings similar to word tokens in NLP.

Although ViTs have shown competitive performance, they require gigantic datasets (such as ImageNet-21k) and a significant amount of computational resources to train effectively. This limits them from being useful in real-world applications like medical imaging, where data are often scarce and/or imbalanced, and computational efficiency is very critical.

Data-efficient Image Transformers (DeiT), introduced by Touvron et al. (2021) [27], overcome this limitation by demonstrating that pure-attention models can be well trained on ImageNet-1k without external data in a few days on a single machine. Touvron et al. mention that their reference vision transformer (86M parameters) achieves a top-1 accuracy of 83.1% (single-crop evaluation) on ImageNet with no external data [27].

One of the innovations brought about by `DeiT` is the distillation token (Figure 2.1) which is a learnable vector that interacts with both the image tokens and a teacher model during training. The distillation token enables Transformer-specific knowledge distillation. In this mechanism, the student `DeiT` model learns not only from ground truth labels but also from the teacher network output, which could be a CNN. This architecture for distillation significantly improves performance in data, allowing DeiTs to work as well as, or even better than, ViT-Base models, but at increased speed and reduced training weight, which makes it an appropriate model for medical imaging applications, including both thoracic and skeletal radiographs.

BERT and TinyBERT

Large pre-trained language models like BERT-base [11] have achieved state-of-the-art performance on all NLP tasks. However, due to their high resource consumption and memory footprint, they are still difficult to deploy in real-world settings, especially when integrated into multi-modal systems.

To overcome these challenges, we can use knowledge distillation, which is a technique that compresses these big models into smaller "student" networks that retain much of the performance of the original model. This process creates a lightweight model that mirrors the outputs and internal representations of a big teacher model.

`TinyBERT`, introduced by Jiao et al. (2019) [16], is a distilled BERT explicitly tailored for Transformer-based models. It suggests a new distillation scheme that aligns not only the end predictions but also the in-between layer-wise representations and attention maps of student and teacher during pre-training. The authors show that a 4-layer TinyBERT achieves over 96.8% BERT-base performance on GLUE, but with 7.5× fewer parameters and 9.4× faster inference [16].

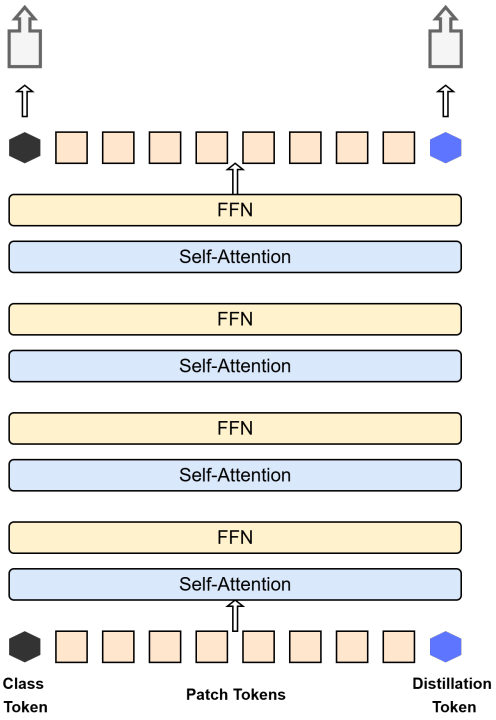


Figure 2.1: Data-efficient image Transformer (DeiT) architecture [1]. The distillation token is what differentiates the `DeiT` architecture from the ViT.

`TinyBERT` follows a two-step approach: pre-training distillation to learn general language knowledge first, followed by task-specific distillation to further refine on downstream tasks. This yields a high-fidelity model with great performance and generalisation, appropriate for use in applications like radiology report interpretation, including findings from chest and skeletal X-rays, where efficiency and accuracy are most important.

Attention

The Transformer architecture was a groundbreaking revelation that replaced the recurrent and convolutional methods with a single, unified mechanism: self-attention [28]. This mechanism computes the representation of each token as a weighted sum of all other tokens, this allows the model to then focus on relevant parts of the input sequence based on learnt relationships.

There are various attention mechanisms that are used in transformers: self-attention, cross-attention, multi-head attention. They all serve different purposes and at different stages of the encoding or decoding process. Self-attention allows a sequence to attend to itself, enabling the model to capture contextual dependencies, such as between tokens in a sentence or patches in an image, within the same input (or the same modality); cross-attention is typically used in decoder modules, where the target sequence (e.g. a translation) attends to the encoder outputs. It can also be adapted for multimodal tasks to let one modality (e.g. text) attend to another (e.g. image).

One of the innovations in the Transformer is the multi-head attention mechanism. The model divides the embedding space into multiple sub-spaces (“heads”), instead of computing attention in a single space. Each head learns varying relationships on its own, it focusses on different parts of the input separately. The subspaces, which are in terms of vectors, are then concatenated and projected back to the model’s original embedding dimension. In this way, the model can be allowed to jointly attend to information at different positions from different representation subspaces, and it can capture richer and more nuanced relationships throughout the sequence [28].

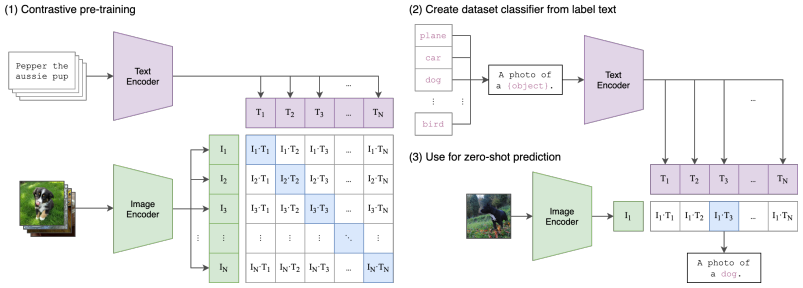


Figure 2.2: The CLIP architecture with separate image and text encoders trained together using a contrastive loss [21].

CLIP

Contrastive Language-Image Pre-training (CLIP) reduces the gap between vision and language by learning joint embeddings for both images and text. This is done by contrasting positively matched image-text pairs against negatively matched ones during training so that the model can project visual and linguistic abstractions without explicit labels or supervision. `CLIP` therefore allows zero-shot transfer learning and can apply a range of downstream tasks without task-specific fine-tuning.

> "We demonstrate that the simple pre-training task of predicting which caption goes with which image is an efficient and scalable way to learn SOTA image representations from scratch on a dataset of 400 million (image, text) pairs...enabling zero-shot transfer of the model to downstream tasks." [21]

The architecture of `CLIP` is composed of two different encoders, an image encoder (typically a Vision Transformer) and a text encoder (a Transformer model), as shown in Figure 2.2. Both encoders are trained with a symmetric contrastive loss that forces the embeddings of the positive image-text pairs to come together and those of the negative pairs to push apart.

This strategy generates highly generalisable representations. `CLIP` demonstrates strong generalisation capabilities in 30+ benchmark tasks, including optical character recognition (OCR), fine-grained image classification, and object detection, to name a few, without any need for additional supervision or retraining.

2.1.3 XAI

As deep learning models become increasingly central to high-stakes applications like medical imaging, interpretability has become essential [26]. Explainable Artificial Intelligence (XAI) strives to provide insight into the decision making of black-box models so that users can better understand, trust, and effectively regulate AI systems [7]. Three of the most important features of XAI in medical imaging are discussed: Gradient-weighted Class Activation Mapping (Grad-CAM), Attention Visualisation for Transformer models, and the White-Box Paradox.

Gradient-weighted Class Activation Mapping (Grad-CAM)

Grad-CAM [23] is a widely used technique for visualising regions in an input image that contribute the most to the prediction of a convolutional neural network (CNN). Through the computation of the gradients of the target class with respect to the feature maps of a convolutional layer, Grad-CAM produces a coarse localisation map of salient regions of the image.

In medical imaging, Grad-CAM has been used to provide visual explanations of model predictions to enable clinicians to understand and confirm AI-driven diagnoses. For example, on medical images like chest or skeletal X-rays, Grad-CAM can identify regions typical of pathologies such as pneumonia, lung nodules, or bone fractures, with visual evidence supporting AI-assisted diagnosis and possibly promoting acceptance of automated systems.

Attention Visualization in Transformer Models

Transformer models, as characterised by their self-attention, have revolutionised the practice of natural language processing and also computer vision applications. Attention visualisation tools allow for the inspection of attention weights between image patches or text tokens to observe where the model is attending to while making the predictions.

While attention maps do offer some insight into what is happening inside the model, it is important to note that large attention weights do not always equal feature importance. Attention visualisation therefore has to be interpreted with caution and, where possible, complemented by other explanation methods in order to gain greater insight into the model's behaviour.

The White-Box Paradox

The "white-box paradox" refers to the seemingly counter-intuitive phenomenon that transparent models or explanations can still lead to user misinterpretation or over-reliance on AI systems. Even when models provide well-defined and understandable explanations, users can have an unjustified level of trust in the AI output, and thus perhaps overlook errors or biases. [2, 5, 6]

This paradox is important for reasons for considering explanations in AI with care. In the medical environment where decisions have significant implications, it is essential that the explanations not only explain the model behaviour, but also accurately depict the reliability and limitations of the AI system. Breaking the white-box paradox is therefore essential to making explainability lead to informed, rather than ill-advised, human oversight.

In this thesis, I develop a multimodal deep learning framework for X-ray diagnostics, focused on chest and appendicular skeletal radiographs, using the publicly available NLMCXR dataset that contains radiographic images and free text radiology reports, as well as clinical data obtained from the Gaetano Pini Orthopedic Institute. In order to effectively use these multimodal data, several techniques covered in this background section are integrated: Data-efficient Vision Transformers (`DeiT`) for image encoding, `TinyBERT` for lighter and faster language understanding, and multiple fusion strategies (including simple concatenation, projection with gating, and cross-attention) to combine visual and textual representations. In addition to that, I also explore `CLIP` as a baseline for contrastive learning to jointly understand vision and language. To ensure interpretability and trustworthiness in medical settings, I incorporate explainability tools like Grad-CAM, while also critically reflecting on their limitations through the lens of the white-box paradox. Together, these components form a robust and transparent pipeline that aims to improve diagnostic accuracy and reliability across a variety of domains in AI-assisted radiology.

2.2 State-Of-The-Art

Deep learning has significantly advanced the field of medical imaging, including chest X-ray (CXR) and appendicular skeletal analysis; it enables diagnostic tools to be automated, more accurate, and scalable. The effort has been directed at developing both unimodal models (vision-only or language-only) as well as multimodal models that exploit the complementarity of images like radiographs and matching clinical text like radiology reports.

2.2.1 Vision-Only Models

Vision-only models have been the predominant solution to automatic medical image interpretation for a while. Convolutional Neural Network (CNN) architectures such as `DenseNet`, `ResNet`, and `EfficientNet` have shown excellent performance in disease classification and localisation in CXRs. An example of a CNN architecture is `CheXNet` [22] which is a 121-layer convolutional network trained on the ChestX-ray14 dataset for pneumonia detection. More recently, Transformer-based models such as Vision Transformers (ViTs) have been introduced, which offer competitive performance by

successfully modelling long-range dependencies and global context in images [8]. One example is the Data-efficient Image Transformer (DeiT), which extends ViTs with competitive performance even in low-data regimes which is quite common in medical imaging. `DeiT` achieves this through distillation-based training, as well as architectural efficiency, and is thus a valid choice for medical applications where datasets are scarce.

Similarly, vision-only models have been used successfully in appendicular skeletal X-ray analysis, particularly for fracture detection using CNNs [17].

2.2.2 Language-Only Models

In parallel with vision advances, large pre-trained language models such as `BERT` and its domain-specific variants (e.g., `ClinicalBERT`, `BioBERT`, `TinyBERT`) have been used in the extraction of structured information from unstructured radiology reports. These models are very effective at capturing contextual and domain-specific semantics and are widely used in applications such as abnormality classification, report summarization, and clinical Named Entity Recognition (NER). In the case of chest or appendicular skeletal X-rays, they can serve as a source of supervision or be integrated into systems that generate or interpret diagnostic reports.

2.2.3 Multimodal Models

Combining image and textual data, multimodal models have shown promise in medical imaging diagnosis, especially in chest X-rays. An example of such models is GE HealthCare's foundation model [15], which is trained on 1.2 million anonymized X-ray images, and it leverages powerful language models to enhance diagnostic capabilities. Open-source initiatives like `ChestX-Transcribe` [24] have combined Swin Transformers for visual feature extraction with `DistilGPT` to produce medical reports, achieving state-of-the-art performance on a wide range of evaluation metrics. Similarly, `CXR-LLaVa` [19] combines large language models with chest X-ray interpretation and aims to mimic the diagnostic expertise of human radiologists. Other than thoracic imaging, multimodal learning has also shown promise in appendicular skeletal X-rays. `BoneCLIP-XGBoost` [25] is an example of such models which combines CLIP-based image features with textual clinical data to assist clinicians in diagnosing bone fractures.

These multimodal approaches reveal the potential for the convergence of vision and language models to improve the accuracy and efficiency of radiographic diagnosis, introducing opportunities for more advanced and accessible healthcare solutions.

Building upon these state-of-the-art advances, this thesis proposes the fusion of vision transformers and language models for both chest and appendicular skeletal X-ray diagnosis. I leverage `DeiT` for efficient image representation and `TinyBERT` for language modelling. Multiple fusion strategies are explored to integrate visual and textual modalities. `CLIP` is also evaluated as a baseline for contrastive learning. To enhance interpretability, the pipeline incorporates Grad-CAM while remaining mindful of challenges like the white-box paradox. This comprehensive multimodal approach aims to improve diagnostic accuracy and enable more interpretable and clinically actionable predictions in X-ray diagnosis.

Chapter 3
Methodology

3.1 Datasets

3.1.1 Open Data (NLMCXR)

The NLMCXR dataset [13, 9], which is an open-access dataset available through the Hugging Face dataset library, was used for this project. It contains 7,430 instances; each instance is composed of three features: a radiology report (text), the path to the image file, and the corresponding chest X-ray image. For the purpose of this project, I only used texts and corresponding images, the paths were unnecessary so they were discarded.

The dataset is already split into training and testing subsets. The training subset has 5,925 instances, while the testing subset has 1,505 instances.

However, the dataset does not contain labels. Since the task of this project is multi-label abnormality classification, labels were necessary and they were extracted manually. This was achieved using a rule-based method that uses a dictionary that maps target labels to associated keywords. I loop through the dataset, and for each text report I check for the presence of these keywords, and at the same time I also check for negations in order to avoid false positives. Using this rule-based method, each instance of the dataset was assigned a list of extracted labels (the findings present in that report and, by extension, the corresponding image or lack thereof). These extracted labels were then encoded into multi-hot label vectors for compatibility with the multi-label classification task.

The set of labels used for classification consists of 14 labels: 13 thoracic conditions that are common on chest radiographs and the absence of any condition. The thoracic conditions are as follows: *Atelectasis, Consolidation, Infiltration, Pneumothorax, Edema, Emphysema, Fibrosis, Effusion, Pneumonia, Pleural Thickening, Cardiomegaly, Nodule/Mass, Hernia*. The lack of a finding was labelled as "*No Finding*". Each instance (image) may be associated with many labels, which shows the multi-label nature of chest X-ray diagnostics.

After extracting the labels, an analysis of the label distribution showed class imbalance. To deal with this, I performed undersampling of the majority class ("*No Finding*"), and I obtained a more balanced label distribution.

Figure 3.1 shows an example of a chest X-ray with the corresponding radiology report.

3.1.2 Clinical Data (Gaetano Pini)

In addition to the open access data obtained from the NLMCXR dataset, I also used the data provided by a doctor at the Gaetano Pini Orthopedic Institute in Milan. The clinical data contains 99 instances, each instance contains two features: images and the corresponding texts. Data were fully anonymised before being delivered to me in order to ensure patient privacy and comply with ethical regulations.

The images are greyscale skeletal radiographs of various regions of the body, including ankles, feet, knees, legs, wrists, hands, pelvis, hips, elbows, and shoulders. The texts are short (unstructured)

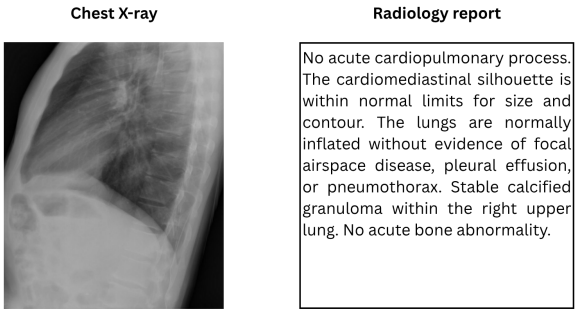


Figure 3.1: Example of a chest X-ray (left) with its corresponding radiology report (right). The image and text report are obtained from the NLMCXR dataset.

free-text reports written by a physician. The texts are written in Italian and consist of 1-3 sentences each in which the observed findings are reported.

The labels were manually assigned by the doctor. The labels were binary and therefore the classification task was also a binary task, with 1 denoting the presence of a fracture and 0 the absence of any abnormalities. The labels were very well balanced with 50 positive cases (fracture) and 49 negative cases (no fracture). The dataset was then split into training, validation, and testing sets with an 80/10/10 split proportion and using stratification to keep the label balance the same in both sets.

An example of a skeletal image and its corresponding text report is shown in figure 3.2.

3.2 Models

This section describes the architecture of the models used in the project. `DeiT` for image encoding, `TinyBERT` for text encoding, and `CLIP` for joint image-text encoding. The section also describes the various fusion techniques used to combine image and text embeddings in multimodal settings. In addition, the section describes the custom classification head across all experiments.

3.2.1 DeiT: Vision Transformer

For encoding the radiographic images, the lightweight variant of Facebook's Vision Transformer, DeiT-tiny (`facebook/deit-tiny-patch16-224`), was used. DeiT [27, 30, 10] is a transformer-based architecture that uses knowledge distillation to make the original Vision Transformer (ViT) architecture more efficient on limited data and computational resources. This "tiny" version is made

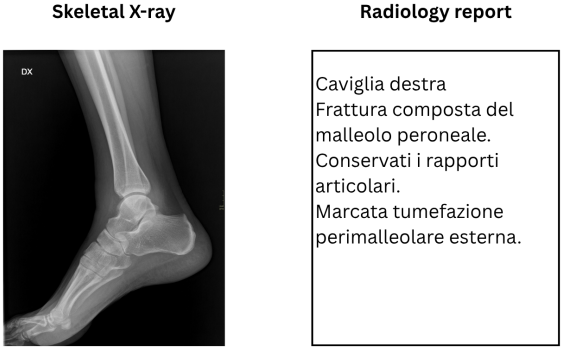


Figure 3.2: Example of a skeletal radiograph (left) and associated radiology report (right) from Gaetano Pini.

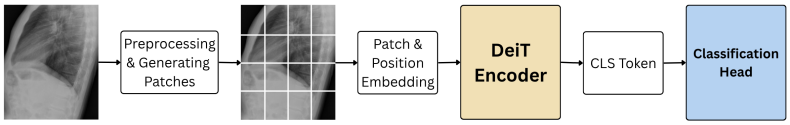


Figure 3.3: A minimal illustration showing an image going through the pipeline for DeiT.

of 12 transformer layers, a hidden size of 192, and 3 attention heads, which makes it significantly lighter (about 5 million parameters) than larger ViT models.

Firstly, the input image is split into 16×16 pixel patches that do not overlap, which results in $(224/16)^2 = 196$ tokens with an input resolution of 224×224. In order to retain spatial information, the patch embeddings are projected onto a linear layer and combined with the positional encodings that have been learnt. A special [CLS] token is added at the beginning of the sequence which will be used for classification. The `DeiT` pipeline is illustrated in Figure 3.3.

In order to reduce training complexity and have a similar backbone across all experiments, the `DeiT` model was implemented as a frozen feature extractor (i.e. all transformer layers were frozen). Only the customised classification head was trained on top of the encoder's output. This also helps to avoid overfitting when working with limited data.

Before encoding, all radiographs were preprocessed in the following way:

- Converted to greyscale (3 channels because the model expects 3 channels)
- Resized to 224×224 pixels to match DeiT's expected input
- Normalized using standard ImageNet preprocessing values (mean and standard deviation)

This setup allowed for efficient extraction of high-level visual features suitable for classification as well as fusion with text embeddings in the multimodal setup.

3.2.2 TinyBERT: Text Encoder

For encoding the radiology reports associated with each X-ray image, the TinyBERT [16] model was used, specifically the `huawei-noah/TinyBERT_General_4L_312D` variant that is openly available through Hugging Face's Transformers library. Through the use of knowledge distillation, this lighter version of BERT was derived from the original BERT-base model, which offers improved inference speed and memory efficiency while maintaining similar performance on downstream tasks. `TinyBERT` is made of 4 transformer layers, a hidden size of 312, and 12 self-attention heads, which makes it suitable for low-resource environments and real-time applications.

Radiology reports were pre-processed using Hugging Face's `AutoTokenizer`, using the same `TinyBERT` model as the backbone. The preprocessing steps were the following:

- Tokenized into word pieces using WordPiece tokenization
- Truncated or padded to a fixed maximum sequence length of 128 tokens
- Converted to input IDs and attention masks

Like in the case of the `DeiT` image encoder, the `TinyBERT` model was implemented as a frozen feature extractor to reduce computational cost and prevent overfitting. Only the custom classification head after the text encoder was trained.

The final embedding corresponding to the [CLS] token was extracted and served as a representation of the radiology report for classification and fusion with image features in the multimodal setting.

3.2.3 CLIP: Joint Image-Text Embeddings

For a separate multimodal experiment, OpenAI's `CLIP` model (`openai/clip-vit-base-patch32`) was used to jointly learn image and text embeddings. `CLIP` (Contrastive Language-Image Pretraining) [21, 18] was developed to learn a common embedding space for images and texts (in natural language) through contrastive learning. During pre-training, `CLIP` learns by associating an image with its corresponding caption by maximising the cosine similarity between the matching image-text pairs and minimising it between mismatched pairs.

CLIP's architecture is made up of two parallel transformer-based encoders:

- A Vision Transformer (ViT-B/32), which splits the images into 32×32 pixel patches, processes them in a similar way to DeiT, and outputs an image embedding.
- A Text Transformer, similar to a lightweight GPT-style, which also outputs a text embedding.

To reduce training complexity, similar to `DeiT` and `TinyBERT`, `CLIP` was used as a frozen encoder; its weights were not updated during training.

In the forward pass of the model:

- Each X-ray image was passed through CLIP's image encoder.
- The associated radiology report was tokenised and also passed through CLIP's text encoder.
- The embeddings that were obtained for the image and text were then concatenated to form a fused multimodal representation.
- The fused embedding was then fed into the custom classification head.

Since the objective of `CLIP` during pre-training involved cosine similarity for retrieval, in this project `CLIP` was repurposed as a general multimodal feature extraction tool. This allowed it to bypass its native contrastive inference mechanism. The extracted image and text features were concatenated and used for the classification tasks.

3.2.4 Fusion

Since the aim of the project is to fuse vision and language representations, different fusion strategies were explored to combine the features extracted from the image and text encoders. In both experiments, fusion was done before classification and all techniques were designed to create a unique representation that captured useful information from both modalities. The strategies fall into two main categories: fusing frozen unimodal encoders (`DeiT` + `TinyBERT`), and joint Image-Text embeddings that were learnt through the use of `CLIP`. The fusion strategies used are shown below.

Concatenation

The simplest and straightforward fusion technique was to directly concatenate the output embeddings from the frozen `DeiT` and `TinyBERT` encoders along the feature dimension. The result was a single vector with the fused embeddings which was then passed to the MLP classification head. This technique does not require training of parameters and maintains both modalities intact.

Projection + Gating

The second technique, as shown in Algorithm 1, first projected both image and text embeddings into a shared dimensional space of size 256. The two projected vectors were then concatenated and passed through a gate to compute a gating vector. This `gate` (g) controlled the balance between the text and image embeddings, combining them as follows:

$$\texttt{fused} = g \cdot \texttt{text} + (1 - g) \cdot \texttt{image}$$

This technique allowed the model to adaptively weigh the impact of text and image features based on their content.

Algorithm 1 Fusion Logic: Projection + Gating

Require: Text embeddings T, Image embeddings I
Ensure: Logits y
$T \leftarrow \text{TextProj}(T)$ // Projected text embeddings
$I \leftarrow \text{ImageProj}(I)$ // Projected image embeddings
$F \leftarrow \text{Concatenate}(T, I)$ // Fused embeddings
$g \leftarrow \text{Gate}(F)$ // Gate values
$F \leftarrow g \cdot T + (1 - g) \cdot I$ // Gated fusion
$F \leftarrow \text{Dropout}(F)$
$y \leftarrow \text{Classifier}(F)$ // Logits
return y

Multi-Head Cross-Attention

The most sophisticated fusion technique was multi-head cross-attention, as shown in Algorithm 2, which was used to allow text embeddings to attend to image embeddings. It worked in the following way:

- Text embeddings were linearly projected to match the dimensions of the image embeddings
- Cross-attention was applied, with 4 heads, using standard attention mechanisms:
  - The image embeddings were the queries
  - The projected text embeddings were the keys and values
- The attention output was an enhanced image embedding which was concatenated with the original text embedding.

This technique allowed the model to focus on image regions conditioned on the text, allowing for a more meaningful interaction between the two modalities.

3.2.5 Classification Head

Across the four chest X-ray experiments, a similar custom classification head, as shown in Algorithm 3, was used for multi-label classification. The classification task was to predict 14 different abnormalities from different input representations: unimodal or multimodal. The architecture of the classification head was the same for every experiment; this was done to ensure consistency and allow fair comparison between experiments.

The classification head was implemented as a small Multi-Layer Perceptron (MLP) consisting of:

- A linear layer projecting from the input embedding dimension to a hidden size of 512 units

Algorithm 2 Fusion Logic: Multi-Head Cross-Attention

Require: Image tokens I_{tok}, Text tokens T_{tok}, Text embeddings T_{emb}
Ensure: Logits y
$(A, \_) \leftarrow \text{CrossAttention}(Q = I_{\text{tok}}, K = T_{\text{tok}}, V = T_{\text{tok}})$ // Cross-attention output
$I_{\text{enh}} \leftarrow A[*, 0]$ // Enhanced image token
$T_{\text{cls}} \leftarrow T_{\text{emb}}[*, 0]$ // [CLS] token from text encoder
$F \leftarrow \text{Concatenate}(T_{\text{cls}}, I_{\text{enh}})$ // Fused embeddings
$F \leftarrow \text{Dropout}(F)$
$y \leftarrow \text{Classifier}(F)$ // Logits
return y

Algorithm 3 Classification Head Forward Pass

Require: Text features t, Image features v
Ensure: Logits y
$f \leftarrow \text{Concatenate}(t, v)$ // Fused features
$h_1 \leftarrow \text{Linear}(f)$ // First linear layer
$h_2 \leftarrow \text{ReLU}(h_1)$ // Activation
$h_3 \leftarrow \text{Dropout}(h_2, p = 0.1)$ // Dropout regularization
$y \leftarrow \text{Linear}(h_3)$ // Final linear layer (logits)
return y

- A `ReLU` activation layer was applied for non-linearity
- A dropout layer (with `dropout_rate=0.1`) for regularization and to prevent overfitting
- A final linear layer projecting from 512 units to 14 output units, one for each label.

During the evaluation, a sigmoid activation function was applied to the model output (logits) to convert them into probabilities in the range of $[0, 1]$. A threshold of 0.5 was then applied to the probabilities to assess whether each abnormality was present (> 0.5) or absent (< 0.5), allowing for multi-label classification.

The input embedding dimension of the classification head depended on the experiment that was being performed:

- In the unimodal cases, it was equal to the CLS token size that was extracted from the final hidden state of the respective encoder:
  - 192 for `DeiT`
  - 312 for `TinyBERT`.
- In the multimodal cases, it depended on the applied fusion strategy:
  - For concatenation, it was simply the sum of the two encoder dimensions ($192 + 312$ or $512 + 512$)
  - For projection and gating, it was the dimensionality of the projection (256)
  - For multi-headed attention, it depended on the dimensionality of the attention output ($312 + 192$)

For appendicular skeletal X-ray experiments, the task was binary classification (presence or absence of a fracture). The classification head in this case was simply a dropout layer (`dropout_rate=0.1`) followed by a linear layer projecting from the input embedding dimension to 1 output unit instead of 14. During the evaluation, sigmoid activation was applied similarly to chest experiments and a threshold of 0.5 was used to make binary predictions.

3.3 Implementation

The project was implemented using ten Jupyter notebooks, the notebooks were divided in two clinical cases: five dedicated to chest X-ray experiments and five to appendicular skeletal X-ray experiments. The structure of the notebooks was similar across the two cases, one notebook was used to prepare the dataset, and the other four were used for the different experimental configurations: DeiT-only, TinyBERT-only, fusion, and CLIP.

In the first notebook of each case, I prepared the dataset. The chest X-ray dataset (NLMCXR) was loaded using the Hugging Face `datasets` library. The instances in this dataset were not labelled, so I used a rule-based approach (Algorithm 4) to extract structured labels and then annotated the dataset with the extracted labels. The labels, for each instance, were in three formats:

- a list of the extracted labels (simply containing the label extracted)
- a vector of ones in the place corresponding to the extracted label and zeros elsewhere (multi-hot vector)
- a binary label (simply indicating the presence/absence of a disease).

The resulting dataset, annotated with the labels, was then saved locally on my disk in Hugging Face format for compatibility with Pytorch's `Dataset` class.

Each of the four experimental notebooks shared a similar end-to-end implementation pipeline consisting of several important steps. The pipeline included:

- loading the dataset that was saved locally using the Hugging Face `dataset` library.
- creating a custom subclass of Pytorch's `Dataset` class to wrap the dataset in a format that is easily compatible with the rest of the implementation. The dataset classes handled images, text, or both modalities and their labels, depending on the experiment.

Data splitting, as described in Section 3.1, was performed before any preprocessing steps to avoid data leakage and overfitting. For preprocessing, the image data were converted to greyscale, resized to 224×224 pixels, transformed into tensors, and normalised using standard ImageNet statistics. For textual data, preprocessing was performed by tokenising using the Hugging Face `AutoTokenizer` with `TinyBERT` as the backbone.

After obtaining the dataset with preprocessed data, the models were defined. Although the architectures of the models are different for the different experiments (as discussed in Chapter 4), the structure and logic used for training are the same. All models were effectively trained using the Hugging Face `Trainer` class. I defined `TrainingArguments` for each experiment to configure the training process. Some of the defined arguments include batch size, learning rate, weight decay, evaluation strategy, and number of epochs.

I defined a custom function to compute the metrics (`compute_metrics`).

Algorithm 4 Label Extraction with Negation Handling

Require: Dataset with text field
Ensure: Extracted labels per dataset entry

```
Convert the input text to lowercase
Split the text into a list of sentences using sentence tokenization
Initialize an empty set matched_labels
for each sentence in the list of sentences do
    for each label, keywords in label_keywords do
        for each pattern in keywords do
            if pattern is found in sentence (using regex) then
                Check if any negation_pattern exists in sentence
                if no negation is found then
                    Add label to matched_labels
                end if
            end if
        end for
    end for
end for
if matched_labels is empty then
    Set extracted_labels to ["No Finding"]
else
    Set extracted_labels to list of matched_labels
end if
return dataset with extracted_labels field updated
```

- In the clinical case of chest X-rays, multi-label classification, the metrics included accuracy, micro F1 score, and micro AUROC.
- In the clinical case of skeletal X-rays, binary classification, the metrics included accuracy, binary F1 score, and AUROC.

After training, I used `Trainer` to save the model and then used the model for evaluation on the test set. Using scikit-learn, I computed and visualised the confusion matrices both for the multi-label case and the binary case to further understand the performance of the model across different abnormality classes or binary outputs.

Using a similar structure across the experiments not only allowed for consistent evaluation of the models but also allowed for fair comparisons between the unimodal and multimodal experiments as well as experiments from different clinical cases.

3.4 Assessment Procedure

The performance of the models was evaluated across both clinical cases: chest X-rays (multi-label classification) and appendicular skeletal X-rays (binary classification). The evaluation was performed using a comprehensive set of metrics, validation strategies, and explainability techniques. I complied, as much as I could, with the CHAMAI checklist, to ensure transparency, robustness, reproducibility, and conformity to medical AI reporting standards. [4, 3]

Multiple evaluation metrics were used to capture different aspects of the performance of the model. The following classification metrics were computed on all the models: Accuracy, Balanced accuracy, Specificity, Sensitivity (Recall), AUROC (area under receiver operating characteristic curve), F1 score, and Matthews correlation coefficient (MCC). Brier score was computed as a calibration metric in order to ensure the model's reliability, i.e. the predicted probability is equal to the true probability. These evaluation metrics were adjusted according to the clinical case. In the chest X-ray case (multi-label) microaveraged F1 score and AUROC were used, as for the case of skeletal X-rays (binary) standard classification metrics were applied (binary average).

The train/validation/test split was performed in each clinical case using an 80/10/10 ratio. In order to ensure consistency and reproducibility, the dataset was divided using a fixed split in all experiments. A dropout layer was added before the last dense layer used for classification in order to reduce overfitting. Cross-validation or hyperparameter optimisation were not performed, since the nature of the project is exploratory.

The evaluation framework was consistent across all models: unimodal models (`DeiT` for vision-only, `TinyBERT` for text-only), fusion models (combining image and text features using concatenation, projection + gating, or multi-head cross-attention), and `CLIP` (pre-trained). The performance of the models were compared using all the metrics, with AUROC the main benchmark. With no specific ablation studies performed, using separate notebooks for each clinical case and modality allowed easy comparison of the contribution of the individual and fused models.

For the chest X-ray clinical case, the radiology reports were used for two purposes:

- **Label Extraction:** Labels were generated from free text reports using a rule-based NLP method. Although no external validation of the quality of the label extraction was done, the keywords were informed by the literature.
- **Model Input:** The reports also served as input for the language models, `TinyBERT` in the language-only modality, `CLIP`, and as part of the fusion models along with the images.

For interpretability and explainability purposes, I computed token-level explainability by calculating the relative token importance by using gradient-based methods similar to Grad-CAM for

images. This helped in understanding the language features that drive the model's decisions and to identify which relevant image regions correspond to key text tokens.

In order to comply with the CHAMAI checklist, the following points were addressed:

- **Transparency:** The preprocessing steps, model architectures, and evaluation metrics are documented in both this thesis and the source code.
- **Code Availability:** The source code comprising all notebooks, Python scripts, and training logs will be made publicly available via a Github repository.
- **Calibration:** The Brier score was calculated to evaluate the calibration/reliability of the models.
- **Explainability:** Explanations were provided through Grad-CAM for the images and the token importance based on the gradient for the texts.
- **Limitations:** The project did not include cross-validation, outlier detection, or hyperparameter optimisation. Data were anonymised, so the demographics of the subjects could not be obtained.

Despite these limitations, the project adheres to the CHAMAI principles in that it provides an end-to-end pipeline that can be reproduced, a strong evaluation framework, and appropriate explainability techniques were applied.

3.5 Tools

To implement this project, several tools and platforms were very useful. The programming language used in the development was Python 3.12.10, using Jupyter Notebooks, which allowed for interaction during experimentation and analysis. The platform to write and execute the code was Microsoft's Visual Studio Code, which was chosen because it is flexible and allows for extended runtimes, which was necessary given the limited computational resources available.

The main deep learning platform used was Pytorch [20] which allows for high-level features like tensor computations and deep neural networks. Aside from this, Hugging Face also played an important role in the project, as it was the source for the datasets as well as the pre-trained models used. The `Transformers` [29] library contributed most to the loading and interface of the model, while the `Trainer` class facilitated the simplification of training procedures and hyperparameter searches.

Other useful libraries included `Scikit-learn` (sklearn) to calculate evaluation metrics such as F1 score, AUROC, and confusion matrices of multi-label and binary tasks, `Matplotlib` and `Seaborn` were used for data visualisation and plots, and Numpy was used for numerical computation and array manipulation. Image preprocessing was achieved using `PIL` and `OpenCV` (cv2), and `pytorch-grad-cam` was used to generate visual explanations of model predictions with Grad-CAM [23]. Basic libraries such as `os` and `random` were also used for file system operations and reproducibility.

The libraries mentioned above were collectively important for the entire pipeline, ranging from data loading and preprocessing to model training, evaluation, and interpretation of results.

Chapter 4
Experimental Setup

This chapter presents how the experiments were set up in order to evaluate the proposed methods in two different clinical cases: chest X-rays and appendicular skeletal X-rays. The chest X-ray dataset (NLMCXR), obtained from an open-access repository (Hugging Face), was chosen because it is readily available and the literature on this modality is extensive, which makes it an appropriate choice for image-text diagnostics. In contrast, the appendicular skeletal X-ray dataset was provided by a medical doctor at the Gaetano Pini Orthopedic Institute in Milan. Its inclusion was useful in demonstrating how the proposed methods would perform on real clinical data, it also allowed for the assessment of the adaptability across domains of the models.

The two datasets have significant differences in language, labelling format, and the focus of diagnosis. The chest X-ray dataset has radiology reports in English and had no labels so required label extraction using a rule-based algorithm to obtain structured labels appropriate for multi-label classification. On the other hand, the skeletal dataset has Italian reports and a separate CSV file with the binary fracture labels. These differences also show how the complexity and structure of the classification tasks varied with multi-label classification for thoracic abnormalities in chest X-rays and binary classification for fracture detection in the case of the skeletal X-rays.

The experimental setup in each clinical case was divided into three parts: a use case that describes how the intended user (radiologist) can use the system in practice, a development pipeline discussing the implementation details in accordance with the CHAMAI checklist, and an experimental pipeline outlining the steps taken in order to evaluate the system's performance.

Despite the differences between the datasets, the experimental pipelines were designed with a parallel structure in order to allow for consistent and methodological comparison between the models. In each clinical case, five Jupyter notebooks were used: one for dataset preparation and four for various experiments: image-only (`DeiT`), text-only (`TinyBERT`), multimodal fusion (using various techniques), and `CLIP`. Although the details inside each notebook were different according to the dataset, the overall flow remained consistent. For example, in the dataset preparation notebook for chest X-rays, it involved label extraction, while for skeletal X-rays it involved file parsing. This parallel structure allowed for a direct and easy comparison between unimodal and multimodal approaches within each clinical case, while also allowing for the assessment of domain adaptability of the models.

4.1 Clinical case 1: Chest X-rays

4.1.1 Use Case

This section provides a description of how a radiologist would interact with the multimodal AI system developed for the classification of abnormalities on chest radiographs. The system should provide diagnostic support to the radiologist in the decision making process, by detecting the presence or absence of abnormalities using both X-ray images and associated radiology reports.

The radiologist would provide two inputs to the system, a chest X-ray image and its corresponding radiology report. The image should be in standard image format and the report should be a plain text file. The model then processes both inputs and outputs a diagnosis in the form of a vector label with indices corresponding to abnormal findings (such as pleural effusion, cardiomegaly) or "No Finding" in the case where no pathology has been detected (see Figure 4.1) along with confidence scores. The output would help the radiologist by serving as a second opinion or a triage tool, especially in low-resource or high-volume environments.

The system interaction, at the moment, takes place in a Jupyter notebook, but in the future, the model could be deployed through a Web-based application. The Web-based application would allow the system to be easily integrated into clinical workflows. The main goal is a tool that would provide radiologists with accurate and timely diagnoses that improve the efficiency and reliability of chest X-ray interpretation.

4.1.2 Development Pipeline

In this section, the development of the AI system for the classification of chest X-ray abnormalities is discussed following the CHAMAI checklist [4, 3]. The checklist ensures transparency, robustness, reproducibility, and is in accordance with medical AI reporting standards.

Clinical Context

The multimodal AI system is designed to be used by radiologists in their routine interpretation of chest X-rays. Its goal is primarily to help clinicians in the diagnosis of chest abnormalities by enriching the process with multimodal AI. The model provides support in diagnostic decisions by providing predictions based on both the X-ray image and the radiology report.

Human Oversight and Intended Use

The system is supposed to be used strictly as a decision support tool. It has been designed with the intention of augmenting, not replacing, the radiologist's judgement. Although no human-in-the-loop validation was implemented in this phase, the system output is supposed to be reviewed by a radiologist in a clinical scenario, where it could be taken into consideration, rejected, or used to prioritise further action.

AI Architecture and Training Pipeline

The model is composed of two frozen encoders: a Vision Transformer (`DeiT`) to encode the images and a language model (`TinyBERT`) to encode the reports. The encoded embeddings are then fused using one of three early fusion techniques: concatenation, projection + gating, or cross-attention (see Figure 4.1). A different multimodal approach was also tried using `CLIP`, which uses its own encoders, and the output is concatenated. A classification head that is simply made of linear layers produces a vector output (length 14, with 1 indicating the presence of that anomaly) across the 14 thoracic abnormalities. The encoders were kept frozen during training, and only the fusion and classification layers were trained.

Model Evaluation and Performance Metrics

The performance of the model was evaluated using multiple metrics such as the F1 score and the AUROC with microaverage, accuracy, Matthews correlation coefficient (MCC), Brier score, specificity, sensitivity (recall), and balanced accuracy. To ensure the reliability of the results, 90% confidence intervals were calculated on all metrics by bootstrapping. These metrics are useful in understanding the performance of the model as well as in comparing between the different multimodal approaches.

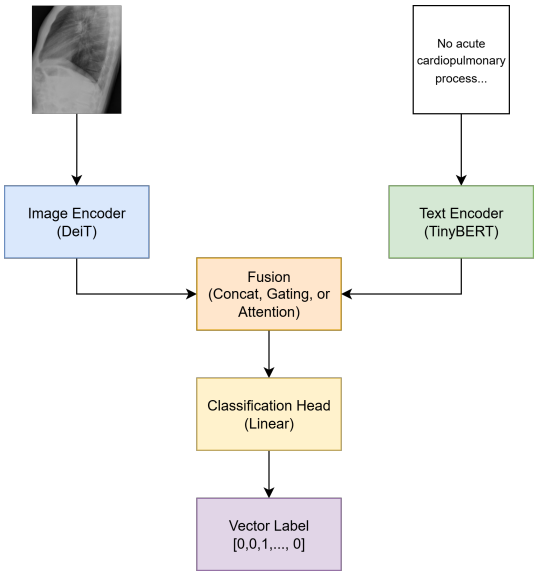


Figure 4.1: Architecture overview of the model, showing a chest X-ray and the associated report as they go through separate image and text encoders, their embeddings fused, pass through the classifier, and the output as a vector of labels.

Auditability and Reproducibility

The model was implemented entirely in Jupyter Notebooks, with a consistent pre-processing structure and fixed random seeds to ensure reproducibility. The NLMCXR dataset, which was used in this case, is publicly accessible through Hugging Face and was already split into training/test sets which were preserved throughout all the experiments. The evaluation setup and model parameters were kept the same across all experiments for comparability.

Interpretability and Explainability

To ensure model transparency, Grad-CAM was used to generate visual explanations to show which regions in the X-ray led to the model's decision. Regarding the reports, gradients were used to show the relative importance of the text tokens. The output of the model was a vector of probabilities for each class which was then thresholded to produce a binary prediction for each class, which supported clinical interpretation.

4.1.3 Experimental Pipeline

The focus of the first clinical case is on the detection of thoracic diseases using the NLMCXR dataset, which is an open access dataset composed of chest X-ray images and their associated radiology reports [9]. The dataset was already divided into train and test sets, with the train set consisting of 5,925 instances and 1,505 instances for the test set. Each instance in the dataset includes a chest radiograph and an associated radiology report in English. The task in this case is multi-label classification because there could be multiple thoracic conditions in a single case. A total of 14 classes were defined: *Atelectasis, Consolidation, Infiltration, Pneumothorax, Edema, Emphysema, Fibrosis, Effusion, Pneumonia, Pleural Thickening, Cardiomegaly, Nodule/Mass, Hernia,* and *No Finding*. The distribution of the labels was inspected and there was a notable label imbalance with the *No Finding* class that comprised approximately two-thirds of the training samples, with most of the other abnormalities under-represented.

Dataset Preparation

The initial processing of the dataset had two main steps: extraction of the labels and balancing of the dataset. Since the NLMCXR dataset did not contain labels, a rule-based label extraction algorithm that uses keyword matching was applied to the radiology reports to obtain a multi-label one-hot vector for the labels. In order to avoid false positives like "no evidence of pleural effusion," negation handling was also explicitly applied in the algorithm. Reports in which no keywords were found were labelled as *No Finding* to indicate the absence of abnormalities.

In order to address the severe class imbalance, undersampling was applied. A unique maximum count was established for all classes and positive instances were selected up to this count or until the samples were exhausted. The result was a dataset with an alleviated class imbalance consisting of 2,291 instances for the train set and 601 instances for the test set. This strategy was used because it prioritised the retention of minority samples while effectively reducing the number of the *No Finding* class.

Due to limited resources and to speed up the training process, the dataset was randomly sampled to obtain training, validation, and testing subsets in a ratio of 80/10/10. This yielded 240 training instances, 30 for validation, and 30 for testing.

Notebook 1: Vision-Only (DeiT)

In this experiment, only images were considered as input to the model. Pre-processing was applied by converting them to greyscale, duplicating the single channel to create a three-channel input since the model expects three channels, resizing them to 224×224, and normalising them using standard ImageNet statistics. The model used was a pre-trained `DeiT-tiny` model [27] that was frozen throughout the training; it served as a feature extractor. For the classification, a custom classification head was added, it consisted of a dropout layer followed by a linear layer that was projecting to 14 output units, one unit for each label. In order to allow for independent prediction of each class, the model was trained using binary cross-entropy with logits as a loss function.

Notebook 2: Language-Only (TinyBERT)

In the text-only experiment the `TinyBERT` model [16] of the Huawei Noah Hugging Face repository was used. The reports were tokenised using the `AutoTokenizer` from Hugging Face, with the maximum sequence length set to 128 tokens. Similarly to the vision model, the text encoder was frozen, and a classification head with a structure identical to the one used for `DeiT` was used. The tokenization procedure produced input ids and attention masks that were used as input to the model. The model was trained using the same binary cross-entropy loss function.

Notebook 3: Multimodal Fusion

For this experiment, the goal was to explore the impact of combining visual and textual features; several fusion strategies were tested to this end. These strategies were as follows:

- **Concatenation:** the embeddings from the frozen `DeiT` and `TinyBERT` encoders were concatenated.
- **Projection + Gating:** the embeddings were first projected into a shared space and combined using a gating vector that has been learnt.
- **Multi-head cross-attention:** this mechanism allowed one modality to attend to features of the other.

An early fusion approach was used in all the fusion strategies, where the embeddings were extracted and fused before being passed to the classification head. The encoders were frozen in all experiments, and only the fusion layers, where applicable, and the classification head were trained.

Notebook 4: CLIP-Based Fusion

The last experiment was carried out using `CLIP ViT-B/32` [21] to learn joint image-text embeddings from image-text pairs. The images and text were preprocessed using the dedicated `CLIP` processor. In order to maintain the same structure throughout all experiments, CLIP's zero-shot capabilities were not relied upon, but rather the embeddings were extracted from the image and text encoders, concatenated, and passed to the custom classification head, which was trained to predict multi-label targets. This setup leveraged a model trained on a broad set of image-text pairs while maintaining a structure that allowed comparison with the other fusion strategies.

Evaluation

In all the experiments, Hugging Face's `Trainer` class was used to train the models. The main evaluation metrics used were accuracy, F1 score (micro), and area under the ROC curve (micro AUROC). Some additional metrics like the Matthews correlation coefficient (MCC), Brier score, specificity, sensitivity (Recall), and balanced accuracy were also calculated to evaluate in a more thorough way the behaviour of the models across classes. The chosen metrics were especially important considering the class imbalance and the multi-label classification task.

4.2 Clinical case 2: Appendicular skeletal X-rays

4.2.1 Use Case

This section provides a description of how a radiologist would interact with the multimodal AI system developed for the detection of fractures on appendicular skeletal radiographs. The system should provide diagnostic support to the radiologist in the decision making process, by detecting the presence or absence of a fracture using both X-ray images and associated radiology reports.

The radiologist would provide two inputs to the system, a skeletal X-ray image and its corresponding radiology report. The image should be in standard image format and the report should be a plain text file. The model then processes both inputs and provides a binary output that indicates diagnosis, whether there is a fracture or not (see Figure 4.2) as well as the confidence score. The output would help the radiologist by providing confirmation of diagnoses or indicating which cases should be given priority, especially in a fast-paced environment like an emergency room.

The system interaction, at the moment, takes place in a Jupyter notebook, but in the future, the model could be deployed through a Web-based application. The Web-based application would allow the system to be easily integrated into clinical workflows. The main goal is a tool that would improve the accuracy and efficacy of the assessment of skeletal trauma for radiologists.

4.2.2 Development Pipeline

Clinical Context

The multimodal AI system is designed for use in an emergency case, in particular in helping clinicians identify fractures in appendicular skeletal X-rays. The system should offer accurate and timely binary fracture predictions based both on the X-ray image and the associated report, thus supporting radiologists or other emergency clinicians.

Human Oversight and Intended Use

The system is supposed to be used strictly as a decision support tool; it is not intended to operate on its own. It has been designed with the intention of augmenting, not replacing, the clinician's judgement. Although no human-in-the-loop validation was implemented in this phase, the labelling of the dataset used in this case was performed by a collaborating physician, which ensured that the

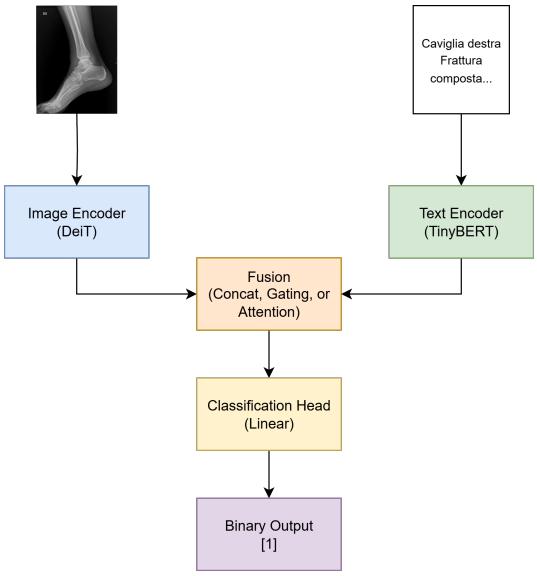


Figure 4.2: Architecture overview of the model, in the skeletal case, showing separate image and text encoders, fusion strategy, final classifier, and binary output.

labels were clinically informed. The output of the model should inform clinician's decisions, not dictate them when the system gets deployed.

AI Architecture and Training Pipeline

In this skeletal case, the core model architecture is the same as in the chest case. The model is composed of two frozen encoders: a Vision Transformer (`DeiT`) to encode the images and a language model (`TinyBERT`) for encoding the reports. The encoded embeddings are then fused using one of three early fusion techniques: concatenation, projection + gating, or cross-attention (see Figure 4.2). A different multimodal approach was also tried using `CLIP` which uses its own encoders and the output is concatenated. A classification head that is simply made up of linear layers produces a binary output indicating the presence or absence of a fracture. The encoders were kept frozen during training, and only the fusion and classification layers were trained just as in the chest case.

Model Evaluation and Performance Metrics

The performance of the model was evaluated using the same metrics used for the chest case, such as F1 score, AUROC, accuracy, Matthews correlation coefficient (MCC), Brier score, specificity, sensitivity (recall), and balanced accuracy. To ensure the reliability of the results, 90% confidence intervals were calculated on all metrics by bootstrapping. These metrics are useful in understanding the performance of the model as well as in comparing between the different multimodal approaches.

Auditability and Reproducibility

The model was implemented entirely in Jupyter Notebooks, with a consistent pre-processing structure and fixed random seeds to ensure reproducibility. Data were provided by a clinician at Gaetano Pini Orthopedic Institute in Milan and had skeletal X-rays and associated reports that were fully anonymised. The data was then randomly split into training, validation, and test subsets using a 80/10/10 split. The evaluation setup and model parameters were kept the same across all experiments for comparability.

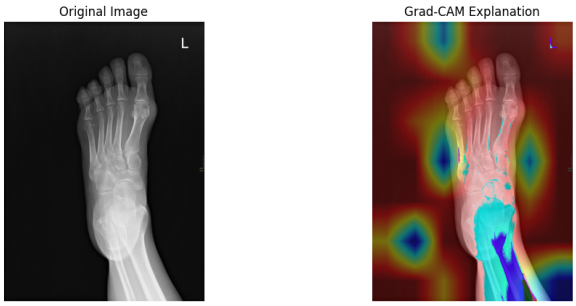


Figure 4.3: Skeletal X-ray with Grad-CAM overlay.

Interpretability and Explainability

To ensure model transparency, Grad-CAM was used to generate visual explanations to show which regions in the X-ray led to the model's decision (see Figure 4.3). Regarding the reports, gradients were used to show the relative importance of the text tokens (see Figure 4.4). The output of the model was the probability of a fracture, which was then thresholded to produce a binary classification decision as the diagnosis.

4.2.3 Experimental Pipeline

In the second clinical case, the focus was on detecting fractures in appendicular skeletal radiographic images. The data consisted of 99 instances that were obtained from a real-world clinical setting and provided by a collaborating medical doctor. Each instance had images of a specific anatomical region (for example, ankle, shoulder, etc.), with a few exceptions which had images of two different regions and also two reports. For the purposes of this project, the anatomical distinctions between the regions were not made; all images were treated uniformly regardless of the region. The dataset was fully anonymised before being used and served as a representative case to test the performance of the models on real clinical data, as well as to test the adaptability of the models across tasks (multi-label or binary), languages (English or Italian), and imaging contexts (thoracic or skeletal).

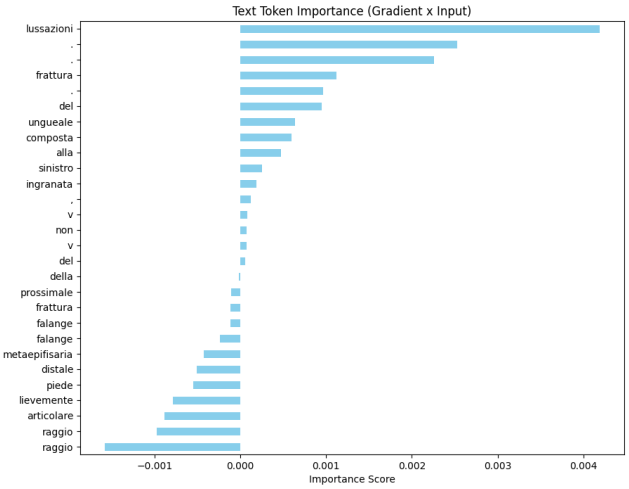


Figure 4.4: Plot showing token importance using gradients.

Dataset Preparation

The dataset was grouped into 99 folders, each containing a unique case (for example, riga 2, riga 3, etc.), corresponding to the rows of an excel file that contained the labels (fracture or no fracture). Each folder consisted of radiographic images and a short radiology report in Italian. These reports were typically started with the body region (such as caviglia destra), followed by diagnostic impressions. No translation or pre-processing was applied to the text reports.

In order to prepare the dataset in a more appropriate format, a dedicated notebook was created for dataset preparation to align and match images, reports, and their labels. The notebook was parsed through an excel file and the folder directory structure to match the contents of each folder with its respective label and report. The dataset produced after this preparation was split into training, validation, and testing sets in a ratio of 80/10/10, which yielded approximately 79 instances for training, 10 instances for validation, and 10 instances for testing. The class distribution of the dataset was balanced, with 50 fracture cases and 49 non-fracture cases.

Notebook 1: Vision-Only (DeiT)

The preprocessing pipeline for the images in this case kept a consistent structure with the chest X-ray pipeline to allow the architectures to be comparable. The following procedures were performed on all radiographic images: conversion to greyscale, expansion to three channels, resizing to 224×224, and normalisation. The `DeiT-tiny` model [27] was used again, pre-trained, and kept frozen during training. A classification head consisting of a dropout layer followed by a linear layer was added to predict the binary label. The loss function used was binary cross-entropy with logits.

Notebook 2: Language-Only (TinyBERT)

Although the reports were written in Italian, the `TinyBERT` model [16] which was trained primarily in English corpora, was used without modification. This was done with the intention of evaluating how well a compact, general-purpose language model could handle generalisation across languages. Similarly to the chest X-ray case, the reports were tokenised using Hugging Face's `AutoTokenizer`, but the maximum sequence length varied in this case and it was 40. The encoder was kept frozen during training and a linear classification layer with dropout was added on top. Surprisingly, the model performed well, eliminating the need for translation or switching to an Italian model or a multilingual model.

Notebook 3: Multimodal Fusion

Like in the chest X-ray case, three fusion strategies using early fusion were tried:

- Concatenating image and text embeddings.
- Projection followed by gating, in which a learnt gate controlled the contribution of each modality.
- Multi-head cross-attention, which allowed for interactions between visual and textual features.

In all cases, the visual and textual encoders were frozen and only the fusion and classification layers were trained. The setup of the architectures maintained consistency to allow for comparison of the different fusion techniques across the two clinical cases.

CLIP-Based Fusion

The last experiment, using `CLIP` [21], was structured the same way as in the previous case. The `ViT-B/32` version of `CLIP` was used to encode both images and text, it produced embeddings that were then concatenated and passed to the classification layer. The official `CLIP` processor was used to handle the preprocessing, and no modifications were applied in order to accommodate the text which was in Italian, CLIP's tokenizer handled it without difficulty.

Evaluation

Given that the classification task was binary and the class distribution was balanced, the evaluation was done using the same metrics: accuracy, F1-score (binary), area under the ROC curve (AUROC) as well as Matthews correlation coefficient (MCC), Brier score, specificity, sensitivity (recall), and balanced accuracy. Although the dataset was small in size, the model pipeline was sufficient to explore the feasibility of multimodal fusion strategies in a different clinical domain.

In this chapter, two clinical cases were presented, chest X-rays and appendicular skeletal X-rays, and they were processed with the intention of using parallel experimental setups and identical model architectures. This choice in the project design was made in order to ensure that the comparison of the models across modalities was fair, as well as to isolate the effects of dataset attributes, such as language (English or Italian), image type (thoracic or skeletal), task (multi-label or binary), and label structure (vector vs integer). Keeping a fixed architecture across both cases allowed the possibility of directly assessing how different clinical domains influenced the performance of models across unimodal and multimodal settings.

For each clinical case, a structured use case, a development pipeline based on the CHAMAI checklist, and an experimental pipeline were described. This setup allowed for in-depth exploration of the behaviour of the models in different clinical contexts. It also enabled the systematic comparison between vision-only, language-only, and multimodal fusion approaches, while also demonstrating the impact of domain shift; such as changing from English to Italian, or from multi-label chest abnormalities to binary fracture classification. The consistency of the experimental design made it easier to assess the generalisability and robustness of the proposed fusion techniques in real-world medical data.

Furthermore, reusing the same architectures and training pipeline was advantageous in terms of simplicity and modularity. After the initial implementation had been implemented and validated in the chest X-ray dataset, it could be extended to the skeletal radiographic task with slight adjustments. We consistently found that multimodal models outperformed unimodal modals across both cases. These findings further validated the central hypothesis of this thesis that the combination of visual and textual information leads to more effective diagnostic performance in medical imaging tasks.

Chapter 5
Results and Discussion

This chapter presents and discusses the results obtained from the experiments performed involving the two clinical use cases of the project: the chest X-ray case and the appendicular skeletal X-ray case. A series of models were trained and evaluated in a consistent manner to ensure that the models were comparable within and between the two clinical cases. The models used include two single-modality models as baseline (DeiT-only for images and TinyBERT-only for text), three fusion strategies (concatenation, projection with gating, and cross-attention) for a multimodal approach, and CLIP as an additional multimodal approach.

Each model was trained in a systematic way and evaluated on a holdout test subset using a comprehensive set of metrics. For the training phase, the training and validation loss, as well as the validation accuracy, F1 score, and AUROC of the best-performing epoch (using F1 score as the evaluation strategy) were reported. The corresponding loss and validation accuracy curves are provided to show the convergence of the models. For the test set, performance is reported in terms of mean accuracy, F1, AUROC, Matthews correlation coefficient (MCC), Brier score, sensitivity, specificity, and balanced accuracy, along with their confidence intervals.

The results are presented in three stages. The first stage is the description of the results in the chest case, followed by the results in the skeletal case. Then, a comparative analysis is provided which highlights the relative strengths of each model across domains. In order to avoid redundancy, loss and accuracy curves are only shown once for each classification task and will be referenced across sections.

In general, from the results two main findings can be obtained: fusion improves performance compared to unimodal approaches, and text-only models (`TinyBERT`) outperform vision-based ones in unimodal settings. These findings are critically analysed in the discussion section.

5.1 Chest Case

In this section, the results obtained from the chest X-ray multi-label classification task are presented. The focus is on the performance of all the six models that were used as mentioned before. In each experiment, the models were trained on a training subset of 240 instances with 30 instances for validation and evaluated on a holdout test set of 30 instances. Given that the nature of the task is multilabel, i.e., most instances are negative for each label, a fixed threshold of 0.5 was applied during evaluation to obtain binary predictions. Evaluation metrics like accuracy, sensitivity, specificity, and balanced accuracy were calculated label-wise and the mean value was obtained, the F1 score was obtained using microaverage and MCC was calculated by flattening all samples and labels. AUROC, like F1, was calculated using microaverage on all labels as well.

The results of the training phase are shown in Table 5.1. The majority of the models took a long time to converge, most likely due to the class imbalance and sparsity of the labels. However, `CLIP` and `TinyBERT` have shown better training and validation performance, with `CLIP` having the lowest

Table 5.1: Training and validation performance on the Chest Case. Values are from the best-performing epoch.

Metric	DeiT-only	TinyBERT-only	Fusion: Concat	Fusion: Gating	Fusion: Attention	CLIP
Train Loss	0.323	0.305	0.263	0.318	0.281	**0.075**
Val Loss	0.315	0.275	0.263	**0.255**	0.282	0.256
Val Acc	0.233	0.233	0.267	0.233	0.267	**0.433**
Val F1	0.209	0.302	0.340	0.311	0.286	**0.576**
Val AUROC	0.565	0.598	0.612	0.593	0.594	**0.744**

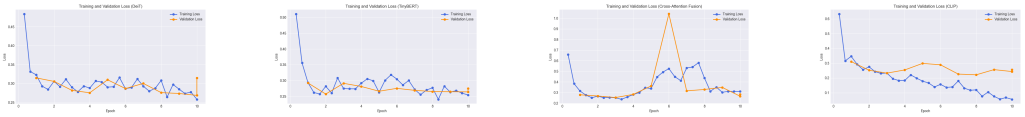
training loss (0.075), the highest validation accuracy(0.433), the highest validation F1 score (0.576), and AUROC (0.744). For custom fusion methods, projection with gating and simple concatenation showed good results; on the other hand, cross-attention underperformed during validation.

The final results obtained from the evaluation on the test set are summarised in Table 5.2. `CLIP` had the highest performance across the main metrics (mean value); these include the AUROC (0.909), Brier score (0.059), and balanced accuracy (0.558), with narrow 90% confidence intervals that indicate that generalisation is stable. The sensitivity (0.147) was also the highest, which makes it the most effective at detecting positive findings. For custom fusion methods, projection with gating slightly outperformed concatenation and cross-attention, in particular, it had higher Specificity and MCC. However, all models had very low sensitivity, which demonstrates the difficulty in predicting rare abnormalities in a multi-label setting that has sparse labels.

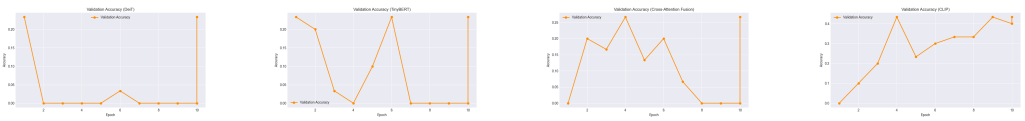
Figure 5.1 shows the corresponding training and validation curves, with complete results on the test set shown in Table 5.2. Overall, these results underline the added value of multimodal learning, in particular with CLIP or well-designed fusion strategies which showed improved performance over the unimodal ones.

Table 5.2: Test performance on the Chest Case with 90% confidence intervals.

Metric	DeiT-only	TinyBERT-only	Fusion: Concat	Fusion: Gating	Fusion: Attention	CLIP
Acc [CI]	**0.528** [0.367–0.667]	0.297 [0.167–0.433]	0.397 [0.233–0.567]	0.296 [0.167–0.433]	0.363 [0.233–0.502]	0.398 [0.267–0.567]
F1 [CI]	**0.483** [0.338–0.635]	0.338 [0.189–0.490]	0.425 [0.271–0.571]	0.386 [0.227–0.542]	0.382 [0.246–0.536]	0.448 [0.298–0.597]
AUROC [CI]	0.776 [0.698–0.851]	0.788 [0.707–0.862]	0.824 [0.744–0.894]	0.811 [0.734–0.892]	0.788 [0.703–0.870]	**0.909** [0.856–0.952]
MCC [CI]	0.441 [0.273–0.617]	0.322 [0.156–0.491]	0.410 [0.239–0.573]	**0.450** [0.307–0.593]	0.354 [0.200–0.519]	0.429 [0.281–0.581]
Brier [CI]	0.066 [0.050–0.083]	0.072 [0.056–0.088]	0.066 [0.051–0.081]	0.064 [0.050–0.078]	0.066 [0.049–0.084]	**0.059** [0.042–0.075]
Sens [CI]	0.071 [0.071–0.071]	0.040 [0.025–0.054]	0.053 [0.040–0.065]	0.040 [0.026–0.054]	0.049 [0.034–0.062]	**0.147** [0.067–0.216]
Spec [CI]	0.929 [0.929–0.929]	0.971 [0.955–0.985]	0.966 [0.948–0.983]	**0.995** [0.986–1.000]	0.951 [0.938–0.967]	0.971 [0.956–0.986]
Bal Acc [CI]	0.500 [0.500–0.500]	0.505 [0.494–0.516]	0.510 [0.497–0.521]	0.518 [0.509–0.526]	0.500 [0.491–0.511]	**0.558** [0.518–0.595]



(a) DeiT Loss **(b)** TinyBERT Loss **(c)** Fusion Loss **(d)** CLIP Loss



(e) DeiT Accuracy **(f)** TinyBERT Accuracy **(g)** Fusion Accuracy **(h)** CLIP Accuracy

Figure 5.1: Training and validation loss and accuracy curves for the Chest case: (a) DeiT Loss, (b) TinyBERT Loss, (c) Fusion Loss, (d) CLIP Loss, (e) DeiT Accuracy, (f) TinyBERT Accuracy, (g) Fusion Accuracy, (h) CLIP Accuracy.

Table 5.3: Training and validation performance on the Skeletal Case. Values are from the best-performing epoch.

Metric	DeiT-only	TinyBERT-only	Fusion: Concat	Fusion: Gating	Fusion: Attention	CLIP
Train Loss	0.879	0.537	0.629	0.588	**0.085**	0.463
Val Loss	1.757	0.512	0.647	0.494	**0.006**	0.512
Val Acc	0.500	0.900	0.700	0.900	**1.000**	0.700
Val F1	0.667	0.889	0.571	0.889	**1.000**	0.571
Val AUROC	0.500	0.900	0.700	0.900	**1.000**	0.700

5.2 Skeletal Case

In this section, the results obtained from the appendicular skeletal X-ray binary classification task are presented. The dataset is made up of 99 instances, split into 79/10/10 for training, validation, and testing, respectively. Unlike the chest case, the task in this case was binary with balanced labels and the clinical reports had considerably shorter text; this contributed to improved performance of the models across all experiments. Similarly to the previous case, a fixed threshold of 0.5 was applied to obtain binary predictions and, for evaluation metrics, standard binary classification.

Most models demonstrated strong behaviour during training as shown in Table 5.3, with cross-attention-based fusion methods achieving perfect validation metrics (F1, AUROC, accuracy = 1.000) and very low validation loss (0.006). However, these results suggest overfitting of the model, most likely due to the small size of the test set as well as the low difficulty of the task compared to the multi-label classification in the other case. The training results were reflected in the evaluation of the test set, where the performance remained perfect, which indicates either overfitting or a very easy classification problem.

The results of the evaluation on the test set are summarised in Table 5.4. The cross-attention fusion method still had perfect results (F1, AUROC, MCC, balanced accuracy = 1), but due to the very small size of the test set and the zero width confidence intervals, this should be interpreted with caution. TinyBERT-only, fusion through projection with gating, and `CLIP` also showed great generalisation abilities, each reaching F1 scores greater than 0.86 and AUROC scores greater than 0.83. However, DeiT-only and fusion using concatenation underperformed showing lower mean scores and wide confidence intervals across all metrics.

`CLIP`, despite the small sample size, maintained strong performance which confirms its robustness and generalisation capabilities across the two clinical domains. However, the instability of some models as shown by the wide confidence intervals and the difficulty in interpreting the perfect scores on a test set with such a small size restricts the conclusions being made with enough certainty.

Figure 5.2 shows the training and validation curves and Table 5.4 presents the results on the test set. These results confirm the findings from the chest case; text-only models perform better than visual-only models, and multimodal fusion approaches (attention or `CLIP`) meaningfully improve performance when regularised and balanced appropriately.

Table 5.4: Test performance on the Skeletal Case with 90% confidence intervals.

Metric	DeiT-only	TinyBERT-only	Fusion: Concat	Fusion: Gating	Fusion: Attention	CLIP
Acc [CI]	0.497 [0.200–0.800]	0.902 [0.700–1.000]	0.603 [0.300–0.800]	0.902 [0.700–1.000]	**1.000** [1.000–1.000]	0.903 [0.700–1.000]
F1 [CI]	0.652 [0.333–0.889]	0.899 [0.667–1.000]	0.646 [0.333–0.889]	0.899 [0.667–1.000]	**0.998** [1.000–1.000]	0.867 [0.600–1.000]
AUROC [CI]	0.438 [0.120–0.793]	**1.000** [1.000–1.000]	0.667 [0.333–1.000]	0.961 [0.833–1.000]	**1.000** [1.000–1.000]	0.837 [0.520–1.000]
MCC [CI]	0.000 [0.000–0.000]	0.818 [0.535–1.000]	0.202 [-0.36–0.667]	0.818 [0.535–1.000]	**0.998** [1.000–1.000]	0.806 [0.509–1.000]
Brier [CI]	0.501 [0.200–0.700]	0.099 [0.000–0.300]	0.402 [0.200–0.700]	0.099 [0.000–0.300]	**0.000** [0.000–0.000]	0.103 [0.000–0.300]
Sens [CI]	**1.000** [1.000–1.000]	**1.000** [1.000–1.000]	0.809 [0.500–1.000]	**1.000** [1.000–1.000]	**1.000** [1.000–1.000]	0.809 [0.500–1.000]
Spec [CI]	0.000 [0.000–0.000]	0.794 [0.500–1.000]	0.393 [0.000–0.800]	0.794 [0.500–1.000]	**1.000** [1.000–1.000]	**1.000** [1.000–1.000]
Bal Acc [CI]	0.500 [0.500–0.500]	0.900 [0.750–1.000]	0.597 [0.333–0.833]	0.900 [0.750–0.938]	**1.000** [1.000–1.000]	0.898 [0.700–1.000]

(a) DeiT Loss **(b)** TinyBERT Loss **(c)** Fusion Loss **(d)** CLIP Loss

(e) DeiT Accuracy **(f)** TinyBERT Accuracy **(g)** Fusion Accuracy **(h)** CLIP Accuracy

Figure 5.2: Training and validation loss and accuracy curves for the Skeletal (Fracture) case: (a) DeiT Loss, (b) TinyBERT Loss, (c) Fusion Loss, (d) CLIP Loss, (e) DeiT Accuracy, (f) TinyBERT Accuracy, (g) Fusion Accuracy, (h) CLIP Accuracy.

5.3 Comparative Discussion

In this section, the results obtained across the two clinical cases are compared and the behaviour of unimodal, fusion-based, and CLIP-based approaches is evaluated. Although the classification tasks were different, multi-label for the chest case and binary for the skeletal case, all models were trained in a similar manner, using frozen backbone encoders, a similar classification head, and binary cross-entropy with logits loss. There were slight variations between experiments, such as the learning rate and the number of output labels between cases.

In both cases, vision-only models underperformed, suggesting that image features alone were not sufficient. On the other hand, language-only models outperformed vision-only models in every case, confirming that clinical text is informative in nature; even when the text is short, as in the second clinical case.

Multimodal approaches showed the most consistent improvements at inference time. `CLIP` was the most robust and generalisable model, with the best performance in the chest case and solid and stable performance in the skeletal case. However, there are certain limitations to speak of in the cases in which it fails. For example, in the skeletal case, the model did not correctly detect a positive fracture case in which the radiology report explicitly mentions the presence of a femoral fracture (“Frattura diametaepifisaria distale del femore...”). Despite this, CLIP’s output was the absence of a fracture with a low confidence score (0.398). This suggests that sometimes the model may undervalue or misinterpret textual information, which highlights the need for better attention and alignment techniques in multimodal models.

Although cross-attention-based fusion approaches achieved perfect test scores in the skeletal case, which should be interpreted with caution because of the zero-width confidence intervals and the small test set. However, gating-based fusion approaches demonstrated solid performance across both domains, suggesting better generalisation than the other custom fusion methods. Fusion with concatenation proved to be effective in the chest case, but was not as effective in the skeletal case.

These results show a clear trend; the fusion of vision and language models consistently improves diagnostic performance, especially when data is limited. Among the various fusion techniques, `CLIP` shows the strongest adaptability across domains, while gating provides a strong trade-off between performance and simplicity. Future work with more resources could help realise the full potential of cross-attention based fusion.

Chapter 6
Conclusions

This thesis investigates the effectiveness of multimodal fusion for radiographic diagnostics using Transformer-based models. The focus of the project was on combining X-ray images and their associated radiology reports to predict abnormalities, using multi-label or binary classification, depending on the clinical case. In this project, two datasets were used: a publicly accessible chest X-ray dataset for multi-label abnormality classification, and hospital appendicular skeletal X-ray data for binary fracture detection.

The findings show that the fusion of image and text modalities improves diagnostic performance compared to unimodal approaches. Language-based models outperform vision-only models, and multimodal techniques `CLIP` in particular showed better generalisation and a more stable training process. These results support the hypothesis that the integration of complementary modalities produces more effective diagnostic predictions.

It is important to note that this thesis should be understood as a scientific investigation of multimodal fusion techniques, rather than the development of a clinically ready deployable system.

6.1 Review of Project Goals

The main goal of this thesis was to evaluate whether combining visual and textual information leads to improved diagnostic performance in medical imaging. The specific objectives were: implementation of unimodal baselines using `DeiT` and `TinyBERT`; implementation and comparison of fusion techniques such as concatenation, gating, attention, and `CLIP`; and evaluation of the adaptability and generalisability of these models across chest and skeletal tasks.

All objectives were achieved successfully. Both unimodal baselines were implemented and were useful as reference points to compare with multimodal approaches. All fusion mechanisms were explored in depth and a systematic comparison of their performance was made in and across the two clinical cases. The results supported the central hypothesis; In the chest case, all the fusion strategies outperformed the unimodal baselines, while in the skeletal case, the language-only model performed better than concatenation, although the other fusion strategies still showed better generalisation. There were some unexpected findings; vision-only models underperformed, and the gating mechanism consistently outperformed concatenation, which is in contradiction to the initial assumptions.

6.2 Limitations

Although the thesis presented the potential of multimodal fusion in radiographic diagnostics, it also presented some clear limitations. The available computational resources were limited, and as a result, the models were kept frozen during training, which limited their capacity for domain adaptation and fine-tuning. The datasets used were relatively small (in the skeletal case) and imbalanced (in the

chest case), reducing the generalisability and clinical applicability of the findings. The model was designed to produce only categorical output as diagnoses given both the radiographic image and the report, which the clinicians considered insufficient. Additionally, the model was not validated on an external independent dataset to test for robustness and hyperparameter optimisation was not performed. Lastly, while some interpretability techniques were considered, a deeper exploration of the models' decision making process was not included in this project.

6.3 Context

This project was carried out with an emphasis on resource efficiency by using lightweight frozen Transformers in order to minimise the environmental impact. The avoidance of fine-tuning hyper-parameters and using pre-trained models shows that meaningful insights can be obtained without immense computational resources. From a societal and ethical point of view, this thesis makes a contribution to diagnostic decision support systems for radiologists, especially in low-resource or fast-paced environments. The project used fully anonymised data, meeting privacy demands, and there is an added emphasis on helping humans rather than replacing them. Although the thesis could potentially be relevant to clinical and enterprise settings, it remains a scientific investigation.

6.4 Future Works

To expand on this thesis, future work could focus on improving the models by fine-tuning or replacing the current models with larger variants or domain-specific models customised for medical imaging and clinical text. The evaluation can also be extended to be more robust by including reliability metrics, like the Expected Calibration Error (ECE) and the Expected Calibration Index (ECI) [14], and fairness along with the expansion of explainability for both image and text inputs. There is also a need for richer outputs; in the future, models can be developed to also generate draft radiology reports from radiographs along with the diagnosis, rather than simple classification tasks. Using the insights from this thesis, the next step would be to develop a multimodal clinical decision support system. This would be done by fusing a Vision Transformer with a language model using projection plus gating. The system's output would be the diagnostic predictions along with confidence scores, then using XAI tools on both inputs to make the outputs more interpretable.

References

[1] Ji-Hyeon Bang, Sung-Wook Park, Jun-Yeong Kim, Jun Park, Jun-Ho Huh, Jung Se Hoon, and Chun-Bo Sim. Ca-cmt: Coordinate attention for optimizing cmt networks. *IEEE Access*, PP:1–1, 01 2023. (Cited on page 10)

[2] Federico Cabitza, Matteo Cameli, Andrea Campagner, Chiara Natali, and Luca Ronzio. Painting the black box white: experimental findings from applying xai to an ecg reading setting, 2022. (Cited on page 12)

[3] Federico Cabitza and Andrea Campagner. The ijmedi checklist for assessment of medical ai, October 2021. (Cited on pages 21 and 24)

[4] Federico Cabitza and Andrea Campagner. The need to separate the wheat from the chaff in medical informatics: Introducing a comprehensive checklist for the (self)-assessment of medical ai studies, 2021. (Cited on pages 21 and 24)

[5] Federico Cabitza, Andrea Campagner, Luca Ronzio, Matteo Cameli, Giulia Elena Mandoli, Maria Concetta Pastore, Luca Maria Sconfienza, Duarte Folgado, Marília Barandas, and Hugo Gamboa. Rams, hounds and white boxes: Investigating human–ai collaboration protocols in medical diagnosis. *Artificial Intelligence in Medicine*, 138:102506, 2023. (Cited on page 12)

[6] Federico Cabitza, Caterina Fregosi, Andrea Campagner, and Chiara Natali. Explanations considered harmful: the impact of misleading explanations on accuracy in hybrid human-ai decision making. In *World conference on explainable artificial intelligence*, pages 255–269. Springer, 2024. (Cited on page 12)

[7] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791, 2021. (Cited on page 11)

[8] Baljinnyam Dayan. Lung disease detection with vision transformers: A comparative study of machine learning methods, 2024. (Cited on page 13)

[9] 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, 07 2015. (Cited on pages 14 and 25)

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. (Cited on page 15)

[11] 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)*, pages 4171–4186, 2019. (Cited on page 9)

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. (Cited on page 9)

[13] Fakhraddin. Nlmcxr dataset. `https://huggingface.co/datasets/Fakhraddin/NLMCXR`, 2024. Accessed: 2025-06-07. (Cited on page 14)

[14] Lorenzo Famiglini, Andrea Campagner, and Federico Cabitza. Towards a rigorous calibration assessment framework: advancements in metrics, methods, and use. In *ECAI 2023*, pages 645–652. IOS Press, 2023. (Cited on page 37)

[15] GE HealthCare. Latest advances in research: Building a multimodal x-ray foundation model. `https://www.gehealthcare.com/insights/article/latest-advances-in-research-building-a-multimodal-xray-foundation-model`, 2024. Accessed: 2025-06-07. (Cited on page 13)

[16] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling BERT for natural language understanding. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 4163–4174. Association for Computational Linguistics, 2020. (Cited on pages 9, 16, 26, and 30)

[17] Nezih Kavak, Rasime Kavak, Bülent Güngörer, Berna Turhan, Sümeyya Kaymak, Evrim Duman, and Serdar Çelik. Detecting pediatric appendicular fractures using artificial intelligence. *Revista da Associação Médica Brasileira*, 70:20240523, 08 2024. (Cited on page 13)

[18] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age, 2019. (Cited on page 17)

[19] Seowoo Lee, Jiwon Youn, Hyungjin Kim, Mansu Kim, and Soon Ho Yoon. Cxr-llava: a multimodal large language model for interpreting chest x-ray images, 2024. (Cited on page 13)

[20] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. (Cited on page 22)

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. (Cited on pages 11, 17, 27, and 30)

[22] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning, 2017. (Cited on page 12)

[23] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. *International Journal of Computer Vision*, 128(2):336–359, October 2019. (Cited on pages 11 and 22)

[24] Prateek Singh and Sudhakar Singh. Chestx-transcribe: a multimodal transformer for automated radiology report generation from chest x-rays. *Frontiers in Digital Health*, Volume 7 - 2025, 2025. (Cited on page 13)

[25] Zhihao Su, Yuhuai Zhou, Jianqun Zhou, Hanhua Cao, and Huanping Zhang. Boneclip-xgboost: A multimodal approach for bone fracture diagnosis. *IEEE Access*, 12:173325–173337, 2024. (Cited on page 13)

[26] Subhashis Suara, Aayush Jha, Pratik Sinha, and Arif Ahmed Sekh. *Is Grad-CAM Explainable in Medical Images?*, page 124–135. Springer Nature Switzerland, 2024. (Cited on page 11)

[27] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. (Cited on pages 9, 15, 26, and 30)

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. (Cited on page 10)

[29] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020. (Cited on page 22)

[30] Bichen Wu, Chenfeng Xu, Xiaoliang Dai, Alvin Wan, Peizhao Zhang, Zhicheng Yan, Masayoshi Tomizuka, Joseph Gonzalez, Kurt Keutzer, and Peter Vajda. Visual transformers: Token-based image representation and processing for computer vision, 2020. (Cited on page 15)