Clinically-aligned Multi-modal Chest X-ray Classification

Phillip Sloan PHILLIP.SLOAN@BRISTOL.AC.UK
Edwin Simpson EDWIN.SIMPSON@BRISTOL.AC.UK
Majid Mirmehdi MAJID.MIRMEHDI@BRISTOL.AC.UK
University of Bristol, United Kingdom

Abstract

Radiology is essential to modern healthcare, yet rising demand and staffing shortages continue to pose major challenges. Recent advances in artificial intelligence have the potential to support radiologists and help address these challenges. Given its widespread use and clinical importance, chest X-ray classification is well suited to augment radiologists workflows. However, most existing approaches rely solely on single-view, image-level inputs, ignoring the structured clinical information and multi-image studies available at the time of reporting. In this work, we introduce CaMCheX, a multi-modal transformer-based framework that aligns multi-view chest X-ray studies with structured clinical data to better reflect how clinicians make diagnostic decisions. Our architecture employs view-specific ConvNeXt encoders for frontal and lateral chest radiographs, whose features are fused with clinical indications, history and vital signs using a transformer fusion module. This design enables the model to generate context-aware representations that mirror the reasoning in clinical practice. Our results exceed the state of the art for both the original MIMIC-CXR dataset and the more recent CXR-LT benchmarks, and highlight the value of clinically grounded multimodal alignment for advancing chest X-ray classification.

Data and Code Availability The primary dataset used in this study was MIMIC-CXR (Johnson et al., 2019) with its CXR-LT 2023/2024 extensions (Holste et al., 2024; Lin et al., 2025). For pretraining, we incorporated additional public datasets, including CheXpert (Irvin et al., 2019a), ChestXray14 (Wang et al., 2017), and VinDR (Nguyen et al., 2022), while MIMIC-IV-ED (Johnson et al., 2023) was leveraged to provide vital signs. Code is available here: https://github.com/phillipSloan/CaMCheX.

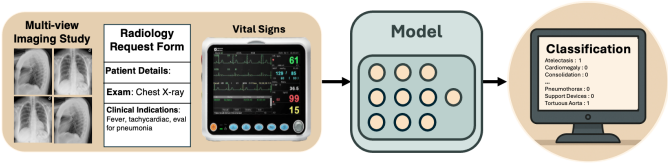

Figure 1: **Conceptual visualisation of CaMCheX's Approach** – CaMCheX mirrors clinical decision-making by incorporating imaging data with complementary information such as clinical indications from radiology request forms and vital signs from A&E triage, enabling accurate and clinically relevant chest X-ray classification.

Institutional Review Board (IRB) This work used public, deidentified imaging datasets with no direct human subject involvement; as such, IRB approval was not required.

1. Introduction

Radiology plays a vital role in modern healthcare, functioning as a key diagnostic tool for identifying diseases, tracking treatment response, and informing patient outcomes. Among radiological procedures, the chest X-ray is the most frequently performed examination and is crucial for the early detection of serious conditions such as pneumonia, lung cancer, and congestive heart failure. However, the increasing gap between radiology workforce capacity and service demand is a growing concern, with nearly all (99%) clinical directors expressing worries about its impact on the safety and quality of care (Royal College of Radiologists, 2023). Fatigue-related errors are an additional concern, potentially compromising diagnostic accuracy (Stec et al., 2018). In light of these challenges, there is an urgent need for innovation,

© 2025 P. Sloan, E. Simpson & M. Mirmehdi.

and AI-based approaches to chest X-ray classification have been identified as a means to augment radiologists' reporting (Royal College of Radiologists, 2018), with multimodal approaches highlighted as essential to maximise impact (Banerji et al., 2025).

Medical image classification presents a distinct set of challenges that set it apart from conventional image classification tasks. For instance, chest X-ray studies often comprise multiple views, including frontal (posteroanterior or anteroposterior) and lateral (left or right) projections, each offering distinct but complementary diagnostic insights. Despite this inherent multi-view structure, the majority of classification methods, such as those proposed in Nguyen-Mau et al. (2023); Jeong et al. (2023); Park et al. (2023); Wang et al. (2024); Dai et al. (2024); Wang et al. (2025), approach the task at single image level, treating each chest X-ray independently. While this has lead to measurable progress, it overlooks the clinical reality that radiologists interpret typically at the study level, where multiple views are reviewed together. Ignoring such anatomical and pathological context across views at study-level may result in missed findings and inaccurate diagnoses.

Medical imaging tasks such as stroke outcome prediction Samak et al. (2022) and stroke lesion segmentation Samak (2025) integrate spatial information across multiple planes or slices. These volumetric methods share conceptual parallels with the multi-view setting, as both combine multiple observations from the same patient. The distinction is that multi-view approaches operate on 2D chest X-rays rather than 3D volumes. Images may include frontal and lateral projections, which provide spatial context, or multiple images from the same perspective, each offering complementary diagnostic information.

Another limitation in current chest X-ray classification methods is the underuse of complementary clinical information by most existing approaches (Kim, 2023; Nguyen-Mau et al., 2023; Jeong et al., 2023). Clinical indications are readily available at the time of reporting and have been shown to enhance diagnostic accuracy in radiological practice (Castillo et al., 2021), and vital signs provide essential physiological context (Candel et al., 2022; Rohmetra et al., 2023), offering cues that can inform image interpretation. For example, an elevated temperature may prompt radiologists to focus more closely on features suggestive of pneumonia. Such structured data have been successfully incorporated into deep learning models in other medical domains to support outcome prediction (Samak et al., 2023; Srivastav et al., 2024). Although Jacenków et al. (2022) incorporate clinical indications, their method remains limited to image-level classification, without study-level, multi-view reasoning or the inclusion of vital signs. To our knowledge, no existing chest X-ray classification approach jointly models both multi-view radiographs and structured clinical information, as found in datasets such as MIMIC-CXR (Johnson et al., 2019) and challenge benchmarks like CXR-LT 2023 (Holste et al., 2024) and CXR-LT 2024 (Lin et al., 2025).

Other long-standing challenges, such as class imbalance and the multi-label nature of chest X-ray datasets, are well recognised in the field (Kim, 2023; Holste et al., 2024). These are commonly addressed using custom loss functions (Kim, 2023; Park et al., 2023) or model ensembling (Nguyen-Mau et al., 2023; Jeong et al., 2023; Hong et al., 2023). We adopt a weighted asymmetric loss function to account for these factors, however, our primary focus is on enhancing multi-view integration and aligning model design more closely with clinical reasoning.

In this paper, we address these limitations by introducing our CaMCheX (Clinical-aligned Multi-modal Chest X-ray), a model that simultaneously integrates clinical indications and vital sign data with multi-view radiological images at the study level, advancing multimodal learning for chest X-ray classification.

CaMCheX incorporates a multi-view image backbone with separate, view-specific ConvNeXt encoders that independently process frontal and lateral images, enabling each encoder to specialise in the distinct characteristics of its respective view and extract complementary features. These radiographic features are then integrated with structured clinical data through a transformer-based fusion module, facilitating context-informed fusion that improves the accuracy and robustness of classification. Figure 1 presents a conceptual overview of our approach.

The design of our framework is grounded in feedback from five clinicians who confirmed they routinely review clinical indications and observations alongside available imaging. Our contributions can be summarised as follows: (i) We propose CaMCheX which integrates multi-view chest X-ray images with structured clinical data, including clinical indications and vital signs, in a manner that more closely reflects the clinical process followed by clinicians (Banerji et al., 2025; Nensa, 2025). (ii) The proposed method is capable of handling any number of frontal and lateral images using separate, view-specific encoders, and

conditionally integrates clinical indications and vital signs when available, enabling robust performance across incomplete or heterogeneous inputs such as those in the MIMIC-CXR dataset. (iii) We perform extensive experiments and ablations to evaluate CaMCheX against state of the art models (Kim, 2023; Li et al., 2024a), CXR-LT 2023 (Holste et al., 2024), CXR-LT 2024 Lin et al. (2025) and MIMIC-CXR (Johnson et al., 2019) label sets, reporting improvements of 20.4% in mean average precision (mAP) and 6.6% in area under the receiver operating characteristic curve (AUROC) for the CXR-LT 2023 benchmark and an improvement of 12.8% in AUROC for the original MIMIC-CXR labels.

2. Related Works

Chest X-ray classification is the automated process of identifying and labeling findings in chest radiographs using machine learning models (Chehade et al., 2024). It is typically formulated as a multi-label classification task, where a single image may exhibit multiple conditions such as pneumonia, cardiomegaly, or pleural effusion. This task supports clinical decision-making by assisting radiologists in detecting pathologies more efficiently and consistently. We focus our review on the under-utilisation of multi-view imaging (Rubin et al., 2018; Zhu and Feng, 2021; Kim, 2023; Lin et al., 2025) and on models that integrate multimodal information into chest X-ray classification (Malik and Anees, 2024; Ketabi et al., 2023; Jacenków et al., 2022). Several of the approaches that we discuss were developed for the CXR-LT 2023 (Holste et al., 2024) and CXR-LT 2024 Lin et al. (2025) challenges. While 2023 entries have separate publications, the 2024 submissions have not appeared independently and are therefore referenced by the challenge summary paper (Lin et al., 2025).

2.1. Multi-view Chest X-ray Classification

Although most chest X-ray classification methods generate predictions at the image level (Nguyen-Mau et al., 2023; Jeong et al., 2023; Park et al., 2023; Hong et al., 2023; Verma, 2023; Yamagishi and Hanaoka, 2023; Kim et al., 2023; Seo et al., 2023; Wang et al., 2024; Dai et al., 2024; Li et al., 2024a; Wang et al., 2025; Lin et al., 2025), this approach departs from clinical practice, where diagnoses are made at the study level by reviewing multiple views. However, several works have begun to explore multi-view approaches. Rubin et al. (2018), Zhu and Feng (2021), and Agostini et al. (2024) implemented architectures to handle frontal and lateral images separately, a limitation of these works is that the architecture adopts a fixed structure and assumes that a frontal and lateral x-ray are available and their designs do not accommodate multiple images from the same view. These assumptions are not reflective of the clinical landscape, or of MIMIC-CXR (Johnson et al., 2019). To address this limitation, competitors of CXR-LT 2023 (Kim, 2023) and CXR-2024 (Lin et al., 2025) developed multi-view classification frameworks to aggregate features across views. While some of these approaches Kim (2023) are invariant to the number and view of the input images, they employ shared encoders for frontal and lateral views which may limit the model's ability to capture the distinct anatomical and contextual features specific to each perspective. Our proposed CaMCheX model builds upon these approaches by employing separate ConvNeXt encoders for frontal and lateral images, enabling the extraction of fine-grained, view-specific features.

2.2. Multi-modal Chest Diagnostics

In medical image classification, multimodality often refers to the combination of imaging types such as CT, MRI, and X-ray (Li et al., 2024b). Other additional clinical data are also increasingly used (Samak et al., 2023, 2025; Srivastav et al., 2024), however such additional modalities are sometimes not available at the time of initial presentation, while clinical indications and triage data are routinely available alongside the chest X-ray in an emergency setting, making them more immediately accessible for diagnostic support. However, to our knowledge, no existing approach combines study-level chest X-ray representations with both clinical indications and vital signs for classification. Several recent works have explored multimodal approaches for chest X-ray classification. Malik and Anees (2024) proposed a multimodal framework combining chest X-rays, CT scans, and recordings of cough sounds. While innovative, this diverges from clinical practice, where reporting clinicians typically lack direct patient contact and thus have no access to auditory cues. Ketabi et al. (2023) introduced a model incorporating chest X-rays, radiology reports, and eye-tracking data to improve classification and explainability. Although this demonstrates the value of textual information, it relies on radiology reports, which are produced post

hoc and are unavailable at the point of initial review. Jacenków et al. (2022) and Shurrab et al. (2024) included some clinical features but their models are limited to image-level predictions and do not incorporate contextual inputs such as vital signs, which offer further diagnostic value. Khader et al. (2023) developed a transformer-based architecture that integrates chest X-ray images with patient vital signs to improve classification performance. However, their approach remains limited to image-level predictions and does not incorporate clinical indications.

The proposed CaMCheX integrates multi-view radiographs with structured clinical data, including clinical indications/history from request forms and vital signs which would be recorded at triage. This reflects real clinical workflows Banerji et al. (2025), where radiologists consider both imaging and context at the study level. CaMCheX enhances classification by aligning with how decisions are made in practice.

3. Methodology

Our proposed CaMCheX model integrates multi-view chest radiographs with structured clinical data. The training follows a two-stage approach. First, we fine-tune a ConvNeXt (Liu et al., 2022) feature extractor for image level chest X-ray classification using noisy student training (Xie et al., 2020), establishing strong visual backbones. In the second stage, this fine-tuned ConvNeXt is incorporated into CaMCheX to encode images as part of a study-level model that fuses radiographic features with clinical indications and vital signs observations.

3.1. Stage 1: Fine-tuning ConvNeXt with Noisy Student Training

In the first stage of our pipeline, we fine-tune a ConvNeXt model (Liu et al., 2022), using image-level chest X-ray classification and noisy student training. This prepares the model to act as a general purpose image feature extractor, producing representations across all projections before it is further adapted to view-specific encoding during the second stage training of CaMCheX. By training on all views, we enable the model to serve as a versatile image feature extractor, capable of producing representations irrespective of the projection before it is fine-tuned further to specific views during Stage 2.

Following Kim (2023), we modify the ConvNeXt architecture to better accommodate the long-tailed, multi-class, multi-label nature of chest X-ray classification by replacing the standard global average pooling and linear classification head with an ML-Decoder (Ridnik et al., 2023), which allows the model to attend to diverse features when identifying different pathologies.

The process can be more formally stated as follows. Given a chest X-ray $\mathbf{x} \in \mathbb{R}^{C \times H \times W}$, where C denotes the number of channels and $H \times W$ the spatial resolution, the ConvNeXt model produces a feature representation $\mathbf{z} \in \mathbb{R}^{C' \times H' \times W'}$. Here, C', H', and W' correspond to the number of output feature channels and the reduced spatial dimensions resulting from convolutional processing. These feature maps are then passed to the MLDecoder classification head, which aggregates the spatial features and generates a prediction vector $\hat{\mathbf{y}} \in \mathbb{R}^{Q}$, where $Q = 26$ corresponds to the number of diagnostic labels defined by the CXR-LT label dataset. These predictions are trained and evaluated against the ground-truth labels $\mathbf{y}$ provided in the CXR-LT dataset (Holste et al., 2024).

To further enhance the quality of the extracted features, we adopt noisy student training (Xie et al., 2020), a method that leverages additional data to improve the initial training process. A teacher model is first trained on the available CXR-LT 2023 labels (Holste et al., 2024) and is subsequently used to generate pseudo-labels for the CheXpert (Irvin et al., 2019a), ChestX-ray14 (Wang et al., 2017) and VinDR (Nguyen et al., 2022) datasets, which are partly labelled, as discussed in Section 4.1. A student model is subsequently trained on the merged set of labelled and pseudo-labelled data which then becomes the teacher for the next iteration, allowing for iterative refinement. In accordance with Xie et al. (2020), we perform four such iterations, beyond which improvements become marginal.

From the jointly trained model, we then train two dedicated encoders, E_F and E_L, on frontal and lateral chest X-rays, respectively. Each is fine-tuned on its corresponding subset of the MIMIC-CXR dataset, enabling the model to adapt its general visual representations to the distinct anatomical and contextual features present in each view. This view-specific training supports more accurate and granular feature extraction, which is essential for effective multi-view classification within the CaMCheX architecture.

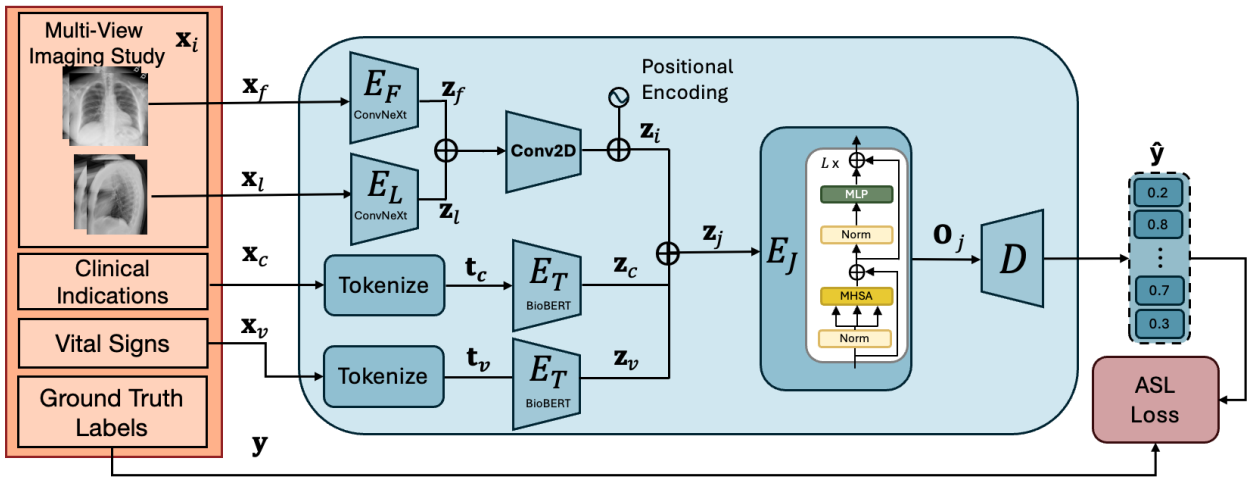

Figure 2: Overview of the proposed CaMCheX architecture (Stage 2). The model processes a set of multi-view images and its associated clinical indications and vital observations to classify chest abnormalities. Features extracted from each modality are distinguished using positional and segment embeddings, concatenated, and fused via a transformer-based fusion module E_j. The resulting joint representation is then passed to a classification head D to produce multi-label predictions $\hat{\mathbf{y}}$.

3.2. Stage 2: Multi-modal, Multi-view Feature Extraction

Figure 2 illustrates the architecture of CaMCheX. Following initial training at the image level, the second phase of our approach performs study-level classification by aggregating multiple chest X-rays per study. We deploy the encoders E_F and E_L trained in Stage 1, while for our textual encoder, E_T, we use BioBERT (Lee et al., 2020), a BERT-based architecture pretrained on various medical texts. This domain-specific pretraining enables BioBERT to capture medical terminology and contextual nuances more effectively than standard BERT models, making it well-suited for extracting clinically relevant features from textual indications.

Consider a set of N frontal (f) and lateral (l) X-ray images $\mathbf{x}_i \in \mathbb{R}^{N\times C\times H\times W}$, split into $\mathbf{x}_f \in \mathbb{R}^{f\times C\times H\times W}$ and $\mathbf{x}_l \in \mathbb{R}^{l\times C\times H\times W}$ respectively, based on view metadata, which is available in MIMIC-CXR (Johnson et al., 2019). Each subset is processed by its respective ConvNeXt feature extractor, yielding $\mathbf{z}_f = E_F(\mathbf{x}_f)$ and $\mathbf{z}_l = E_L(\mathbf{x}_l)$, where $\mathbf{z}_f \in \mathbb{R}^{f\times C'\times H'\times W'}$ and $\mathbf{z}_l \in \mathbb{R}^{l\times C'\times H'\times W'}$. The resulting features are merged into a joint tensor which is then passed through a shared convolutional layer to produce spatially reduced maps $\mathbf{z}_i \in \mathbb{R}^{N\times C'\times H''\times W''}$. Positional encodings are then added to preserve spatial information within the transformer.

As N varies between studies (within MIMIC-CXR, $1 \leq N \leq 11$, with $0 \leq f \leq 10$ and $0 \leq l \leq 5$), we cap the number of images per study at $N = 4$ for efficiency and memory constraints. For those with more than four, a random subset is sampled each epoch to preserve computational efficiency, which only impacts 191 out of 227,827 studies in MIMIC-CXR.

The structured clinical inputs, clinical indications ($\mathbf{x}_c$) and vital signs ($\mathbf{x}_v$), are independently processed by the BioBERT encoder E_T. Each input can be represented as a token sequence, $\mathbf{x}_a = [w_1, w_2, \ldots, w_S]$, which is tokend to $\mathbf{t}_a = [t_1, t_2, \ldots, t_n]$ and passed through the encoder to obtain contextual representations: $\mathbf{z}_a = E_T(\mathbf{t}_a) \in \mathbb{R}^{n\times d}$. To integrate with the image features, the first token embedding $\mathbf{z}_a^{\text{cls}} \in \mathbb{R}^d$ is extracted and reshaped to match the spatial dimensions of the image feature maps, yielding $\mathbf{z}_a^{\text{spatial}} \in \mathbb{R}^{1\times C'\times H''\times W''}$. This process generates the features $\mathbf{z}_c$ and $\mathbf{z}_v$ for clinical indications and vital signs, respectively.

Once encoded segment embeddings are applied to allow the transformer to be able to distinguish between the generated features $\mathbf{z}_i$, $\mathbf{z}_c$ and $\mathbf{z}_v$, they are then concatenated along the sequence dimension to form $\mathbf{z}_j \in \mathbb{R}^{(N+2)\times C'\times H''\times W''}$. The tensor is then

passed through a transformer encoder E_J, producing the joint output, $\mathbf{O}_j$, which our classification head, D, decodes to generate a prediction vector $\hat{\mathbf{y}} \in \mathbb{R}^Q$. For studies missing clinical indications or vital signs, placeholders are used during both training and inference which allows the model to maintain a consistent input structure while preventing the introduction of synthetic or biased values. Example inputs to the model can be seen in Table A1.

4. Experiments

Within this work, MIMIC-CXR (Johnson et al., 2019) with the CXR-LT 2023 (Holste et al., 2024) labels serves as the primary dataset to fine-tune and evaluate our model, while several additional datasets are used during noisy student training to increase the number of training samples available. We also make use of the MIMIC-IV (Johnson et al., 2023) dataset to extract patient vital signs from the available electronic hospital records. While we perform supervd training with the CXR-LT labels, we report evaluatation results for CaMCheX on both the MIMIC-CXR and CXR-LT label sets.

4.1. Datasets

We use several datasets in this work. **MIMIC-CXR**(Johnson et al., 2019) is the largest available chest x-ray classification dataset, containing 377,110 radiographic images from 227,835 radiology studies conducted on 65,379 individuals presenting to the Beth Israel Deaconess Medical Center Emergency Department. Each study has an associated radiology report containing the following sections: indication, technique, comparison, findings, and impression. The clinical indications used in our model are extracted from the indication section of the corresponding radiology reports using the NLP extraction tool provided by Johnson et al. (2019), while diagnostic labels for training and evaluation are derived from the same reports using the CheXpert rule-based labeler (Irvin et al., 2019a), which assigns 14 categories corresponding to common thoracic diseases and findings.

Holste et al. (2024) introduce a set of MIMIC-CXR classification labels comprising of 26 categories and later expanded to 40 labels in 2024 Lin et al. (2025). As a newly established benchmark, it plays a key role in defining the current state of the art in chest X-ray image classification. The CXR-LT 2024 benchmark defines three tasks, but we focus on task 2, long tailed classification on a small, manually annotated test set of the same 26 labels as the CXR-LT 2023 benchmark. **MIMIC-IV-ED** (Johnson et al., 2023) is an emergency department dataset containing structured data from 441,561 visits by 322,872 patients again at Beth Israel Deaconess Medical Center. It includes triage observations such as temperature, heart rate, respiration rate and blood pressure, and can be linked to MIMIC-CXR via shared identifiers such as study-id and admission times. The overlap of patients between MIMIC-CXR and MIMIC-IV-ED is $61,856$.

We incorporate three additional chest X-ray datasets into our training pipeline. **CheXpert** (Irvin et al., 2019a) consists of 224,316 chest X-ray images from 65,240 patients, labelled using a rule-based labeler developed for a medical image classification competition (Irvin et al., 2019b) to extract 14 classification labels which can be considered positive, negative or uncertain. For Noisy Student training (Xie et al., 2020) we retain positive and negative labels and treat uncertain or unmentioned cases as unlabelled. Building on CheXpert, CheXpert Plus (Chambon et al., 2024) released the radiology reports for each study in (Irvin et al., 2019a), although, at present, these reports are only available for the training and development sets. **NIH ChestX-ray14** (Wang et al., 2017) contains 112,120 frontal chest X-ray images from 30,805 patients, each annotated with up to 14 disease labels. **VinDr-CXR** (Nguyen et al., 2022) includes 67,914 chest X-ray images collected from two hospitals in Vietnam, annotated with 15 abnormality labels, with Nodule and Mass findings merged into a single Nodule/Mass class for label consistency with the CXR-LT 2023 labels. For all of these three datasets, we apply soft pseudo-labeling to any classes that do not overlap with the CXR-LT 2023 label set. Neither NIH ChestX-ray14 (Wang et al., 2017) nor VinDr-CXR (Nguyen et al., 2022) include associated clinical information.

4.2. Implementation details

Model implementation was developed in PyTorch and PyTorch Lightning, leveraging pretrained backbones from the timm and transformers libraries. Images were resized to 1024×1024. Training was performed using the AdamW optimizer with a learning rate of 3×10^{-5} and a weight decay of 1×10^{-4} across all phases. To combat the long tail and label sparsity, we adopted a weighted asymmetric loss (ASL) (Ben-Baruch et al., 2021) function. A cosine-annealing

learning-rate schedule with a 5% linear warm-up phase was used, with a batch size of 16. Exponential Moving Average (EMA) of model weights was used to stabilise learning. Training was performed on a single GH200 Superchip with 96Gb of vRAM. We repeat each experiment 3 times with different random seeds, reporting the mean in our results.

In line with the CXR-LT benchmarks (Holste et al., 2024; Lin et al., 2025), we evaluate CaMCheX with commonly used metrics, mAP and AUC-ROC (macro). The primary evaluation metric is mAP since AUC-ROC can be disproportionately influenced due to long tailed class imbalance (Holste et al., 2024) In contrast, mAP measures performance across decision thresholds and does not degrade under class imbalance (Holste et al., 2024). For benchmarking against prior work, we also report results on the original 14-label MIMIC-CXR label dataset, where AUROC remains the standard metric of comparison.

4.3. Results

Table 1 compares the performance of our proposed CaMCheX model against a range of recent approaches. On the CXR-LT 2023 benchmark our model achieves a mAP of 0.576 and an AUROC of 0.916, outperforming all competing methods by a substantial margin across both evaluation metrics. Compared to CheXFusion (Kim, 2023), which previously led the benchmark with an mAP of 0.372 and AUROC of 0.850, CaMCheX demonstrates a relative improvement of 20.4% and 6.6% in mAP and AUROC respectively. Notably, several other top entries in the challenge, including those of Nguyen-Mau et al. (2023), Jeong et al. (2023), and Park et al. (2023) also rely on ConvNeXt backbones but do not match our performance, underscoring the benefit of our multimodal design rather than architectural choice alone.

On Task 2 of the CXR-LT 2024 benchmark, our model achieves a mAP of 0.725 and an AUROC of 0.904, establishing a new state of the art. The strongest competing entries for the CXR-LT 2024 benchmark are Team C and E (Lin et al., 2025) who scored highest on mAP and AUROC respectively. Team C employed a multimodal ensemble approach that combined multi-view, multi-scale image alignment with contrastive image-language pretraining, reaching an mAP of 0.526 and AUROC of 0.833. Team E used a foundational model, the DINOv2 model (Oquab et al., 2023), along with multi-view, multi-resolution ensembling with a self-distillation loss, achieving 0.511 for mAP and 0.836 for AUROC. Compared to these approaches, CaMCheX delivers an improvement of 19.9% in mAP and 6.8% in AUROC, underscoring the effectiveness of our multimodal design over the ensemble approaches within the CXR-LT 2024 competition.

When comparing against methods evaluated on the MIMIC-CXR labels, we see an improvement of 12.8% over the state of the art (Jacenków et al., 2022), demonstrating that our approach is generalisable beyond the CXR-LT setting, reinforcing the robustness of our multimodal design.

As seen in Table 2, for the CXR-LT labels CaMCheX achieves the highest mAP on head classes, scoring 0.786, which significantly surpasses all baseline methods. The next best, CheXFusion (Kim, 2023), achieves 0.499, highlighting CaMCheX 's strong ability to accurately detect the most common pathologies. For medium-frequency (body) classes, CaMCheXagain substantially outperforms other methods, achieving a mAP of 0.550. This result is particularly noteworthy, as it more than doubles the score of CheXFusion, which records 0.246. While some existing methods show strong head-class performance, they typically struggle in the medium range, whereas CaMCheX demonstrates clear robustness. Finally, in the tail group, CaMCheX achieves a mAP of 0.408, representing a marked improvement over previous methods such as Jeong et al. (2023), who achieve an mAP of 0.246. This gain on rare pathologies underscores CaMCheX 's ability to generalise effectively across the entire distribution, maintaining strong performance even on the most infrequent conditions.

Table A2 presents per-pathology results on the CXR-LT 2023 benchmark, grouped by class frequency. Overall performance follows the expected long-tail trend, with head classes achieving the highest mean AP (0.786) due to their greater representation and more distinctive visual features. Body classes exhibit moderate performance (mean AP = 0.550), indicating that the model generalizes well even when fewer samples are available. Tail classes, while the most challenging (mean AP = 0.408), still maintain strong discriminative capability despite limited training data. Notably, Pleural Other remains a difficult category, reflecting both its rarity and the inherent ambiguity of its label. These results demonstrate that, although performance naturally correlates with class frequency, our model performs robustly across all groups. This consistent performance across the head, body, and tail distributions high-

Model	Dataset	Year	Resolution	Architecture	mAP	AUROC
Seo et al. (2023)	CXR-LT 2023	2023	1024	ResNet101	0.279	0.782
Kim et al. (2023)	CXR-LT 2023	2023	448	TResNet50	0.328	0.822
Yamagishi and Hanaoka (2023)	CXR-LT 2023	2023	224, 512	EfficientNetV2	0.330	0.826
Verma (2023)	CXR-LT 2023	2023	448	ResNeXt101 & DenseNet161	0.339	0.830
Hong et al. (2023)	CXR-LT 2023	2023	512	ResNet50	0.349	0.836
Park et al. (2023)	CXR-LT 2023	2023	1024	ConvNeXt	0.351	0.838
Jeong et al. (2023)	CXR-LT 2023	2023	448	ConvNeXt	0.354	0.838
Nguyen-Mau et al. (2023)	CXR-LT 2023	2023	512/768	EfficientNetV2-S & ConvNeXt	0.354	0.836
CheXFusion (Kim, 2023)	CXR-LT 2023	2023	1024	ConvNeXt & Transformer	0.372	0.850
Ours (CaMCheX)	CXR-LT 2023	2025	1024	ConvNeXt & Transformer	**0.576**	**0.916**
Team F (Lin et al., 2025)	CXR-LT 2024	2024	384, 512	ConvNeXt & MaxViT-T	0.509	0.829
Team E (Lin et al., 2025)	CXR-LT 2024	2024	336, 448, 512	ViT-L	0.511	0.836
Team A (Lin et al., 2025)	CXR-LT 2024	2024	224, 384	ConvNeXt	0.519	0.834
Team C (Lin et al., 2025)	CXR-LT 2024	2024	1024	ConvNeXt & EfficientNet	0.526	0.833
Ours (CaMCheX)	CXR-LT 2024	2025	1024	ConvNeXt & Transformer	**0.725**	**0.904**
DualNet* (Rubin et al., 2018)	MIMIC-CXR	2018	-	DenseNet-121	-	0.721
MMVM* (Agostini et al., 2024)	MIMIC-CXR	2024	224	ResNet VAE	-	0.733
Shurrab et al. (2024)	MIMIC-CXR	2024	-	Siamese ViTs	-	0.751
Yao et al. (2025)	MIMIC-CXR	2025	224	Instruction ViT	-	0.782
MBRANet (Li et al., 2024b)	MIMIC-CXR	2024	224	ResNet50	-	0.801
MMBT (Jacenków et al., 2022)	MIMIC-CXR	2022	224	ResNet-50 & BERT	-	0.806
Ours (CaMCheX)	MIMIC-CXR	2025	1024	ConvNeXt & Transformer	**0.669**	**0.934**

Table 1: Quantitative evaluation with state of the art approaches for the MIMIC-CXR, CXR-LT 2023, and CXR-LT 2024 (Task 2) benchmarks. Best method's result is in bold, and second best is underlined. Results marked with * are reported on different MIMIC-CXR splits.

lights that leveraging complementary information effectively enhances classification performance for both frequent and rare pathologies.

Altogether, these results demonstrate that CaMCheX delivers consistently strong performance across all frequency groups, outperforming the state of the art in both common and rare pathology detection.

4.4. Ablation

Table 3 presents an ablation study evaluating the contribution of each component on the CXR-LT 2023. Starting from a single-view baseline, we observe an improvement when moving to a multi-view setup, with a 3.6% increase in mAP and 1.7% increase in AUROC. This demonstrates the benefit of study-level classification over image-level, which provides complementary anatomical information. Examining the contribution of the non-imaging modalities in isolation, clinical indications achieve moderate performance of 0.268 mAP and 0.745 AUROC, suggesting clinical indications provide useful context but they are insufficient on their own to capture the full complexity of thoracic observations. Vital signs performs marginally above chance when used in isolation, with a mAP of 0.128 and an AUROC of 0.580, suggesting that measures such as heart rate are, on their own, weak predictors of thoracic findings. When integrated with imaging features, however, both modalities enrich the model's generated features.

It can be seen that incorporating clinical information into the multi-view model improves performance, improving mAP by 16.5% in mAP and 4.4% in AUROC, indicating that the inclusion of contextual clinical information supports more accurate predictions. Integrating vital signs achieves the highest performance, with CaMCheX achieving a further improvement of 2.5% in mAP and 1.3% in AUROC. These results demonstrate that each modality, multi-view images, clinical indications and vital signs contribute to the models capabilities.

5. Conclusions and Future Work

We introduce CaMCheX, a multimodal framework for chest X-ray classification that integrates multi-view radiographs with structured clinical data, including clinical indications and vital signs. The design of our framework is grounded in feedback from five clinicians, who confirmed that they routinely review clinical context alongside both frontal and lateral chest radiographs as part of their diagnostic

Model	Year	mAP		
		Head (>10%)	Body (1–10%)	Tail (<1%)
Seo et al. (2023)	2023	0.420	0.154	0.086
Kim et al. (2023)	2023	0.461	0.195	0.195
Yamagishi and Hanaoka (2023)	2023	0.460	0.195	0.216
Verma (2023)	2023	0.476	0.210	0.179
Hong et al. (2023)	2023	0.474	0.218	0.243
Park et al. (2023)	2023	0.480	0.221	0.220
Jeong et al. (2023)	2023	0.477	0.226	0.246
Nguyen-Mau et al. (2023)	2023	0.482	0.226	0.227
CheXFusion (Kim, 2023)	2023	0.499	0.246	0.242
Ours (CaMCheX)	-	**0.786**	**0.550**	**0.408**

Table 2: Per-frequency-group mAP results (Head, Medium, Tail) for the CXR-LT 2023 benchmark. Best result in bold, second is underlined.

Model	Images	Multi-View (MV)	Clinical Indications (CI)	Vital Signs (VS)	mAP	AUROC
Single-View	✓				0.346	0.837
CI only			✓		0.268	0.745
VS only				✓	0.128	0.580
Multi-View	✓	✓			0.392	0.859
MV + CI	✓	✓	✓		0.551	0.903
CaMCheX	✓	✓	✓	✓	**0.576**	**0.916**

Table 3: Ablation study illustrating the contribution of each modality to overall performance on the CXR-LT 2023 benchmark.

workflow. By aligning image features with structured clinical information through a transformer-based module, CaMCheX more closely mirrors real-world diagnostic reasoning, where radiologists integrate diverse information sources to build a clinical picture that informs their decisions (Nensa, 2025). Our model employs view-specific encoders and a context-aware fusion mechanism, enabling it to capture both anatomical and physiological cues.

Evaluated on the MIMIC-CXR and CXR-LT benchmarks, CaMCheX achieves state of the art performance for both tasks. In addition to strong overall results and transferability across datasets, it demonstrates robust generalisation across head, body, and tail classes. These findings highlight the value of clinically grounded, multimodal models in advancing automated chest X-ray interpretation. At present, clinical information such as indications and other demographics is not available for the test set of CheXpert Plus (Irvin et al., 2019a) preventing direct validation of CaMCheX against methods that adopt CheXpert as a baseline. If this becomes available it will allow a more rigorous evaluation against further models.

Image resolutions have been shown to influence performance in chest X-ray classification (Kim, 2023) and exploring the impact of resolution within our multimodal framework represents a valuable direction for future work. While CaMCheX incorporates clinical information typically available to radiologists at the time of interpretation, such as vital signs and clinical indications, other context that informs diagnostic reasoning, such as patient demographics, were not available for this work due to dataset scope. Prior imaging is available within MIMIC-CXR and offers a valuable opportunity for longitudinal analysis, which we identify as an important direction for future work.

Looking ahead, the CaMCheX encoder can be integrated into automated radiology report generation (ARRG) systems, a field growing rapidly (Sloan et al., 2024). As this area expands, grounding generative models in clinically validated, multimodal features is increasingly important. Our classification results offer a robust foundation for future systems to generate accurate, context-aware, and clinically meaningful reports, aligning these ARRG systems with multimodal clinical information.

Acknowledgement

The 1st Author wishes to acknowledge and thank the financial support of the UKRI (Grant ref EP/S022937/1) and the University of Bristol. The authors acknowledge the use of resources provided by the Isambard-AI National AI Research Resource (AIRR). Isambard-AI is operated by the University of Bristol and is funded by the UK Government's Department for Science, Innovation and Technology (DSIT) via UK Research and Innovation; and the Science and Technology Facilities Council [ST/AIRR/I-A-I/1023].

References

Andrea Agostini, Daphné Chopard, Yang Meng, Norbert Fortin, Babak Shahbaba, Stephan Mandt, Thomas M. Sutter, and Julia E. Vogt. Weakly-supervised multimodal learning on mimic-cxr, 2024. URL https://arxiv.org/abs/2411.10356.

Christopher R. S. Banerji, Aroon Bhardwaj Shah, Ben Dabson, Tapabrata Chakraborti, Vicky Hellon, Chris Harbron, and Ben D. MacArthur. Clinicians must participate in the development of multimodal ai. *eClinicalMedicine*, 84, 2025. ISSN 2589-5370. doi: 10.1016/j.eclinm.2025.103252. URL https://doi.org/10.1016/j.eclinm.2025.103252. Published online June 2025.

E. Ben-Baruch, T. Ridnik, N. Zamir, A. Noy, I. Friedman, M. Protter, and L. Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 82–91. IEEE, 2021.

B. G. J. Candel, R. Duijzer, M. I. Gaakeer, E. ter Avest, Ö. Sir, H. Lameijer, R. Hessels, R. Reijnen, E. W. van Zwet, E. de Jonge, and B. de Groot. The association between vital signs and clinical outcomes in emergency department patients of different age categories. *Emergency Medicine Journal*, 39(12):903–911, 2022.

C. Castillo, T. Steffens, L. Sim, and L. Caffery. The effect of clinical information on radiology reporting: A systematic review. *Journal of Medical Radiation Sciences*, 68(1):60–74, Mar 2021. doi: 10.1002/jmrs.424. Epub 2020 Sep 1.

P. Chambon, J.-B. Delbrouck, T. Sounack, S.-C. Huang, Z. Chen, M. Varma, S. Q. H. Truong, C. T. Chuong, and C. P. Langlotz. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. *arXiv preprint arXiv:2405.19538*, 2024.

A. H. Chehade, N. Abdallah, J.-M. Marion, M. Hatt, M. Oueidat, and P. Chauvet. A systematic review: Classification of lung diseases from chest x-ray images using deep learning algorithms. *SN Computer Science*, 5:405, 2024. doi: 10.1007/s42979-024-02751-2. URL https://doi.org/10.1007/s42979-024-02751-2.

T. Dai, R. Zhang, F. Hong, J. Yao, Y. Zhang, and Y. Wang. Unichest: Conquer-and-divide pre-training for multi-source chest x-ray classification. *IEEE Transactions on Medical Imaging*, 43(8): 2901–2912, 2024. doi: 10.1109/TMI.2024.3381123.

G. Holste, Y. Zhou, S. Wang, A. Jaiswal, M. Lin, S. Zhuge, Y. Yang, D. Kim, T.-H. Nguyen-Mau, M.-T. Tran, J. Jeong, W. Park, J. Ryu, F. Hong, A. Verma, Y. Yamagishi, C. Kim, H. Seo, M. Kang, L. A. Celi, and Y. Peng. Towards long-tailed, multi-label disease classification from chest x-ray: Overview of the cxr-lt challenge. *Medical Image Analysis*, 97:103224, 2024.

F. Hong, T. Dai, J. Yao, Y. Zhang, and Y. Wang. Bag of tricks for long-tailed multi-label classification on chest x-rays. *arXiv preprint arXiv:2308.08853*, 2023.

J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019a.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea, Hayden Chute, Hershel Mehta, Tony Duan, Daisy Cheng, Daniel Y. Fung, Brian H. Jones, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert

comparison. https://stanfordmlgroup.github.io/competitions/chexpert/, 2019b. Accessed: 15 May 2025.

G. Jacenków, A. Q. O'Neil, and S. A. Tsaftaris. Indication as prior knowledge for multimodal disease classification in chest radiographs with transformers. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2022. doi: 10.1109/ISBI52829.2022.9761567.

J. Jeong, B. Jeoun, Y. Park, and B. Han. An optimized ensemble framework for multi-label classification on long-tailed chest x-ray data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2739–2746, Paris, France, 2023. IEEE.

A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C. Deng, R. G. Mark, and S. Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(317), 2019.

A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C. Deng, R. G. Mark, and S. Horng. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10(1), 2023. doi: 10.1038/s41597-022-01899-x.

S. Ketabi, P. Agnihotri, H. Zakeri, K. Namdar, and F. Khalvati. Multimodal learning for improving performance and explainability of chest x-ray classification. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023 Workshops*, pages 107–116, Cham, 2023. Springer Nature Switzerland.

F. Khader, G. Müller-Franzes, T. Wang, T. Han, S. Tayebi Arasteh, C. Haarburger, J. Stegmaier, K. Bressem, C. Kuhl, S. Nebelung, J. N. Kather, and D. Truhn. Multimodal deep learning for integrating chest radiographs and clinical parameters: A case for transformers. *Radiology*, 309(1):e230806, oct 2023. doi: 10.1148/radiol.230806.

C. Kim, G. Kim, S. Yang, H. Kim, S. Lee, and H. Cho. Chest x-ray feature pyramid sum model with diseased area data augmentation method. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2757–2766, Paris, France, 2023. IEEE.

D. Kim. Chexfusion: Effective fusion of multi-view features using transformers for long-tailed chest x-ray classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2702–2710, Paris, France, 2023. IEEE.

J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020. doi: 10.1093/bioinformatics/btz682.

D. Li, H. Huo, S. Jiao, X. Sun, and S. Chen. Automated thorax disease diagnosis using multi-branch residual attention network. *Scientific Reports*, 14 (1):11865, 2024a. doi: 10.1038/s41598-024-62813-6. URL https://doi.org/10.1038/s41598-024-62813-6.

Y. Li, M. El Habib Daho, P-H. Conze, R. Zeghlache, H. Le Boité, R. Tadayoni, B. Cochener, M. Lamard, and G. Quellec. A review of deep learning-based information fusion techniques for multimodal medical image classification. *Computers in Biology and Medicine*, 177:108635, 2024b. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2024.108635. URL https://www.sciencedirect.com/science/article/pii/S0010482524007200.

Mingquan Lin, Gregory Holste, Song Wang, Yiliang Zhou, Yishu Wei, Imon Banerjee, Pengyi Chen, Tianjie Dai, Yuexi Du, Nicha C. Dvornek, Yuyan Ge, Zuwei Guo, Shouhei Hanaoka, Dongkyun Kim, Pablo Messina, Yang Lu, Denis Parra, Donghyun Son, Álvaro Soto, Aisha Urooj, René Vidal, Yosuke Yamagishi, Pingkun Yan, Zefan Yang, Ruichi Zhang, Yang Zhou, Leo Anthony Celi, Ronald M. Summers, Zhiyong Lu, Hao Chen, Adam Flanders, George Shih, Zhangyang Wang, and Yifan Peng. Cxr-lt 2024: A miccai challenge on long-tailed, multi-label, and zero-shot disease classification from chest x-ray. *Medical Image Analysis*, 106:103739, 2025. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2025.103739. URL https://www.sciencedirect.com/science/article/pii/S1361841525002865.

Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Com-*

puter Vision and Pattern Recognition (CVPR), pages 11976–11986, 2022.

Hassaan Malik and Tayyaba Anees. Multi-modal deep learning methods for classification of chest diseases using different medical imaging and cough sounds. *PLOS ONE*, 19(3):e0296352, 2024. doi: 10.1371/journal.pone.0296352.

Felix Nensa. The future of radiology: The path towards multimodal ai and superdiagnostics. *European Journal of Radiology Artificial Intelligence*, 2:100014, 2025. ISSN 3050-5771. doi: https://doi.org/10.1016/j.ejrai.2025.100014. URL https://www.sciencedirect.com/science/article/pii/S305057712500012X.

H. Q. Nguyen, K. Lam, L. T. Le, H. H. Pham, D. Q. Tran, D. B. Nguyen, D. D. Le, C. M. Pham, H. T. T. Tong, D. H. Dinh, C. D. Do, L. T. Doan, C. N. Nguyen, B. T. Nguyen, Q. V. Nguyen, A. D. Hoang, H. N. Phan, A. T. Nguyen, P. H. Ho, D. T. Ngo, N. T. Nguyen, N. T. Nguyen, M. Dao, and V. Vu. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. *Scientific Data*, 9:429, 2022. doi: 10.1038/s41597-022-01558-1.

T. H. Nguyen-Mau, T. L. Huynh, T. D. Le, H. D. Nguyen, and M. T. Tran. Advanced augmentation and ensemble approaches for classifying long-tailed multi-label chest x-rays. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2729–2738, Paris, France, 2023. IEEE.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

W. Park, I. Park, S. Kim, and J. Ryu. Robust asymmetric loss for multi-label long-tailed learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2711–2720, Paris, France, 2023. IEEE.

T. Ridnik, G. Sharir, A. Ben-Cohen, E. Ben-Baruch, and A. Noy. Ml-decoder: Scalable and versatile classification head. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 32–41. IEEE, 2023.

H. Rohmetra, N. Raghunath, P. Narang, V. Chamola, M. Guizani, and N. R. Lakkaniga. Ai-enabled remote monitoring of vital signs for covid-19: Methods, prospects and challenges. *Computing*, 105: 783–809, 2023.

Royal College of Radiologists. The standards for interpretation and reporting of imaging investigations, 2nd ed. https://www.rcr.ac.uk/publication/standards-interpretation-and-reporting-imaging-investigations-second-edition, 2018. [Online; accessed 2025-05-11].

Royal College of Radiologists. Clinical radiology uk workforce census 2022 report. Technical report, The Royal College of Radiologists, June 2023.

J. Rubin, D. Sanghavi, C. Zhao, K. Lee, A. Qadir, and M. Xu-Wilson. Large scale automated reading of frontal and lateral chest x-rays using dual convolutional neural networks. *arXiv preprint*, 2018.

Z. A. Samak. Multi-type stroke lesion segmentation: comparison of single-stage and hierarchical approach. *Medical & Biological Engineering & Computing*, 63:975–986, 2025. doi: 10.1007/s11517-024-03243-4. URL https://doi.org/10.1007/s11517-024-03243-4.

Z. A. Samak, P. Clatworthy, and M. Mirmehdi. Transop: Transformer-based multimodal classification for stroke treatment outcome prediction. In *Proceedings of the IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2023. doi: 10.1109/ISBI53787.2023.10230576.

Z. A. Samak, P. Clatworthy, and M. Mirmehdi. Automatic prediction of stroke treatment outcomes: latest advances and perspectives. *Biomedical Engineering Letters*, 15:467–488, 2025. doi: 10.1007/s13534-025-00462-y.

Zeynel A. Samak, Philip Clatworthy, and Majid Mirmehdi. Fema: Feature matching auto-encoder for predicting ischaemic stroke evolution and treatment outcome. *Computerized Medical Imaging and Graphics*, 99:102089, 2022. ISSN 0895-6111. doi: https://doi.org/10.1016/j.compmedimag.2022.102089. URL https://www.sciencedirect.com/science/article/pii/S0895611122000623.

H. Seo, M. Lee, W. Cheong, H. Yoon, S. Kim, and M. Kang. Enhancing multi-label long-tailed classification on chest x-rays through ml-gcn augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2747–2756, Paris, France, 2023. IEEE.

S. Shurrab, A. Guerra-Manzanares, and F. E. Shamout. Multimodal masked siamese network improves chest x-ray representation learning. *Scientific Reports*, 14:22516, 2024. doi: 10.1038/s41598-024-74043-x. URL https://doi.org/10.1038/s41598-024-74043-x.

P. Sloan, P. Clatworthy, E. Simpson, and M. Mirmehdi. Automated radiology report generation: A review of recent advances. *IEEE Reviews in Biomedical Engineering*, 18:368–387, 2024. doi: 10.1109/RBME.2024.3408456.

S. Srivastav, M. Ranjit, F. Pérez-García, K. Bouzid, S. Bannur, D. C. Castro, A. Schwaighofer, H. Sharma, M. Ilse, V. Salvatelli, S. Bond-Taylor, F. Falck, A. Thieme, H. Richardson, M. P. Lungren, S. L. Hyland, and J. Alvarez-Valle. Maira at rrg24: A specialised large multimodal model for radiology report generation. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing (BioNLP)*, pages 597–602, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.bionlp-1.50. URL https://aclanthology.org/2024.bionlp-1.50.

N. Stec, D. Arje, A. R. Moody, E. A. Krupinski, and P. N. Tyrrell. A systematic review of fatigue in radiology: Is it a problem? *American Journal of Roentgenology*, 210(4):799–806, 2018.

A. Verma. How can we tame the long-tail of chest x-ray datasets?, 2023. URL https://arxiv.org/abs/2309.04293.

B. Wang, H. Pan, A. Aboah, Z. Zhang, E. Keles, D. Torigian, B. Turkbey, E. Krupinski, J. Udupa, and U. Bagci. Gazegnn: A gaze-guided graph neural network for chest x-ray classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2194–2203, January 2024.

M. Wang, Y. He, L. Peng, X. Song, S. Dong, and Y. Gong. Cross-domain invariant feature absorption and domain-specific feature retention for domain incremental chest x-ray classification. *IEEE Transactions on Medical Imaging*, 44(5): 2041–2055, 2025. doi: 10.1109/TMI.2025.3525902.

X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2097–2106. IEEE, 2017.

Q. Xie, M.-T. Luong, E. Hovy, and Q. V. Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10687–10698. IEEE, 2020.

Y. Yamagishi and S. Hanaoka. Effect of stage training for long-tailed multi-label image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2721–2728, Paris, France, 2023. IEEE.

Junjie Yao, Junhao Wang, Zhenxiang Xiao, Xinlin Hao, and Xi Jiang. Intelligent analysis of chest x-ray based on multi-modal instruction tuning. *Meta-Radiology*, 3(3):100172, 2025. ISSN 2950-1628. doi: https://doi.org/10.1016/j.metrad.2025.100172. URL https://www.sciencedirect.com/science/article/pii/S2950162825000402.

X. Zhu and Q. Feng. Mvc-net: Multi-view chest radiograph classification network with deep fusion. In *Proceedings of the IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 554–558. IEEE, 2021. doi: 10.1109/ISBI48211.2021.9434000.

Appendix A. Example Inputs

Study ID	Images	Clinical Indications	Vital Signs (text input)
50000766	2	year old woman s/p right vats mediastinal bx and pericardial effusion pna pna .	Temperature: 97.9 | Heart rate: 109.0 | Respiratory rate: 22.0 | O2 Saturation: 94.0 | Systolic BP: 136.0 | Diastolic BP: 54.0 | Gender: F
50000801	3	year old man with bilateral pleural effusions s/p drainage r/o ptx r/o ptx .	Temperature: 98.8 | Heart rate: 70.0 | Respiratory rate: 18.0 | O2 Saturation: 100.0 | Systolic BP: 99.0 | Diastolic BP: 56.0 | Gender: M
50020872	2	-year-old female with shortness of breath .	Temperature: 101.2 | Heart rate: 115.0 | Respiratory rate: 18.0 | O2 Saturation: 100.0 | Systolic BP: 116.0 | Diastolic BP: 77.0 | Gender: F
50173309	2	year old woman with exertional dyspnea mild hypoxia eval for acute process .	Temperature: 98.5 | Heart rate: 89.0 | Respiratory rate: 18.0 | O2 Saturation: 98.0 | Systolic BP: 127.0 | Diastolic BP: 74.0 | Gender: F
50286142	2	hodgkins lymphoma atypical chest pain evaluation .	Temperature: 99.8 | Heart rate: 130.0 | Respiratory rate: 20.0 | O2 Saturation: 96.0 | Systolic BP: 84.0 | Diastolic BP: 48.0 | Gender: M
50349057	1	-year-old female with history of chest pain left rib pain .	Temperature: 98.3 | Heart rate: 68.0 | Respiratory rate: 16.0 | O2 Saturation: 99.0 | Systolic BP: 135.0 | Diastolic BP: 72.0 | Gender: F

Table A1: Example study-level inputs showing clinical indications and the vital-signs text are represented and processed by our model.

Appendix B. Pathology-Level results

Pathology	**AP**	**AUROC**
Head (>10%)		
Atelectasis	0.6569	0.8489
Cardiomegaly	0.6999	0.8426
Edema	0.8275	0.9507
Lung Opacity	0.6512	0.8152
No Finding	0.7409	0.9398
Pleural Effusion	0.8848	0.9509
Pneumonia	0.8813	0.9451
Support Devices	0.9447	0.9674
Mean (Head)	**0.786**	**0.920**
Body (1-10%)		
Calcification of the Aorta	0.2174	0.9300
Consolidation	0.4987	0.8877
Emphysema	0.2674	0.9274
Enlarged Cardiomediastinum	0.3011	0.7006
Fracture	0.6807	0.9411
Hernia	0.7284	0.9585
Infiltration	0.8976	0.9768
Mass	0.6413	0.9365
Nodule	0.3883	0.8757
Pneumothorax	0.8827	0.9714
Mean (Body)	**0.550**	**0.911**
Tail (<1%)		
Fibrosis	0.4428	0.9575
Lung Lesion	0.7221	0.9715
Pleural Other	0.0537	0.9142
Pleural Thickening	0.1508	0.8906
Pneumomediastinum	0.6579	0.9755
Pneumoperitoneum	0.4677	0.9645
Subcutaneous Emphysema	0.6530	0.9924
Tortuous Aorta	0.1651	0.8760
Mean (Tail)	**0.408**	**0.943**

Table A2: Per-pathology classification results on the **CXR-LT 2023** dataset grouped by Head, Body, and Tail categories depending on their an . MeAverage Precision (mAP) and Area Under the ROC Curve (AUROC) are reported for each pathology, along with group means.