

Multimodal Deep Learning for Chest X-Ray Abnormality Classification and Interpretability through Demographic Feature Integration

Sultan Mesfer Aldossary

Prince Sattam Bin Abdulaziz University

Samia M. Abd-Alhalem

Delta University for Science and Technology

Magda M. Algameel

Prince Sattam Bin Abdulaziz University

Noha E. El-Attar

`noha.ezzat@fci.bu.edu.eg`

Benha University

Research Article

Keywords: ConvNeXt, patient demographics, multi-label classification, chest X-ray, NIH ChestX-ray14, transfer learning, Grad-CAM

Posted Date: August 18th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10284358/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Multimodal Deep Learning for Chest X-Ray Abnormality Classification and Interpretability through Demographic Feature Integration

Sultan Mesfer A Aldossary¹, Samia M. Abd-Alhalem², and Magda M. Algameel³, Noha E. ElAttar^{2,4*}

¹ Department of Computer Engineering and Information, College of Engineering in Wadi Alddawasir, Prince Sattam bin Abdulaziz University, Wadi Al-Dawasir, Saudi Arabia

² Department of Bioinformatics, Faculty of Artificial Intelligence, Delta University for Science and Technology, Gamasa, Dakahlia, Egypt,

³ Department of Nursing, College of Applied Medical Sciences, Prince Sattam bin Abdulaziz University, Wadi Al-Dawasir, Saudi Arabia.

⁴ Department of Information Systems, Faculty of Computers and Artificial Intelligence, Benha University, Benha, Egypt.

Corresponding author: Noha E. AlAttar (e-mail: noha.ezzat@fci.bu.edu.eg).

“The authors extend their appreciation to Prince Sattam bin Abdulaziz University for funding this research work through the project number (PSAU/2025/1/38421).”

Abstract

Chest radiography remains one of the most widely used imaging modalities for the diagnosis and screening of thoracic diseases. However, the increasing volume of radiological examinations and the complexity of identifying multiple coexisting abnormalities continue to challenge clinical workflows. Although deep learning has achieved substantial progress in automated chest X-ray analysis, most existing approaches rely exclusively on image information and do not fully exploit clinically relevant patient metadata. This study proposes a multimodal deep learning framework that integrates chest radiographic images with demographic information, including patient age, sex, and radiographic view position, for multi-label thoracic disease classification. The proposed framework employs a ConvNeXt Large backbone for image feature extraction and a dedicated metadata-processing branch to learn complementary patient-specific representations. To address the severe class imbalance inherent in the NIH ChestX-ray14 dataset, weighted binary cross-entropy loss was incorporated during training. Patient-level data partitioning was enforced throughout the experimental pipeline to prevent information leakage between training, validation, and testing subsets.

The framework was evaluated on the NIH ChestX-ray14 dataset, comprising 112,120 frontal chest radiographs from 30,805 patients spanning fourteen thoracic disease categories. Experimental evaluation yielded a macro-average area under the receiver operating characteristic curve (AUC) of 96.88%. Under the same evaluation protocol, the proposed framework outperformed ResNet-152 (85.02%), DenseNet-121 (87.76%), and DenseNet-169 (87.21%). Ablation analysis further confirmed the positive contribution of both demographic metadata integration and imbalance-aware optimization to overall model performance. Grad-CAM visualizations additionally demonstrated that model predictions were associated with anatomically relevant image regions, providing qualitative evidence supporting model interpretability. The observed performance improvements demonstrate that incorporating demographic metadata alongside image features enhances thoracic disease discrimination, supporting the value of multimodal learning for large-scale chest radiograph analysis.

Keywords: *ConvNeXt; patient demographics; multi-label classification; chest X-ray; NIH ChestX-ray14; transfer learning; Grad-CAM*

1. Introduction

Chest radiography plays a central role in the diagnosis, screening, and follow-up of thoracic diseases owing to its rapid acquisition, low cost, and widespread availability in routine clinical practice [1]. Chest

X-rays are commonly used to assess pulmonary and cardiovascular abnormalities, including pneumonia, pleural effusion, pneumothorax, pulmonary nodules, cardiomegaly, and other thoracic findings that require timely clinical interpretation [1][2]. The large volume of radiographic examinations performed daily in hospitals and diagnostic centers has increased the demand for automated systems that can assist radiologists and improve the consistency of image interpretation [3][4].

Despite its clinical value, chest X-ray interpretation remains challenging. Many thoracic abnormalities present with subtle visual patterns, and several diseases may overlap in appearance or coexist within the same radiograph [5]. Diagnostic interpretation may also vary across readers because of differences in expertise, workload, and fatigue. Recent evidence further suggests that radiology workloads are expected to continue increasing, while artificial intelligence is increasingly being investigated as a tool to support workflow efficiency and clinical decision-making [3][4]. Collectively, these challenges have increased interest in computer-aided diagnosis systems that can support radiologists during large-scale chest radiograph interpretation.

Deep learning has significantly advanced medical image analysis by enabling models to learn hierarchical visual representations directly from imaging data [6]. In chest radiography, convolutional neural networks have achieved strong performance in thoracic abnormality detection, multi-label disease classification, and image interpretation tasks [7][8]. Progress in this area has been accelerated by the release of large public chest X-ray datasets. ChestX-ray8 introduced a hospital-scale dataset for weakly supervised thoracic disease classification [7]. CheXpert provided a large chest radiograph dataset with uncertainty labels extracted from radiology reports [9]. MIMIC-CXR released a large de-identified database of chest radiographs paired with free-text reports [10]. VinDr-CXR further contributed expert radiologist annotations for thoracic abnormality assessment [11]. Several deep learning architectures have been explored for automated chest X-ray interpretation. DenseNet-based models received considerable attention following their use in CheXNet for pneumonia detection and multi-label thoracic disease classification [8]. ResNet and DenseNet variants remain widely adopted baseline models because of their established performance in image recognition and their suitability for transfer learning in medical imaging [12][13]. More recently, ConvNeXt combines the efficiency of convolutional neural networks with several design principles inspired by transformer-based architectures, enabling strong representation learning while maintaining scalability for large-scale image analysis tasks [14]. Such characteristics have motivated the adoption of ConvNeXt in demanding image classification tasks where robust feature representation is essential. Despite recent advances in deep learning-based chest X-ray analysis, important methodological challenges remain unresolved. Many existing systems rely primarily on image information and do not fully exploit patient-specific metadata that is routinely available in clinical datasets. Demographic attributes such as age, sex, and radiographic view position may provide complementary diagnostic information that helps distinguish visually similar abnormalities. These attributes are commonly collected in clinical practice and are available in several public chest radiograph datasets; however, their integration with modern deep learning architectures for large-scale multi-label thoracic disease analysis remains relatively underexplored [15]. In addition, severe class imbalance remains a persistent challenge in multi-label chest X-ray datasets, where frequent findings appear far more often than rare but clinically important abnormalities [7][9]. This imbalance can bias model learning toward common classes and reduce performance for underrepresented diseases.

Interpretability is another important requirement for clinical use. Deep neural networks may produce accurate predictions, but their decision-making process is often difficult to inspect. In radiology, visual explanations are valuable because they allow clinicians to examine whether the model focuses on anatomically plausible regions. Grad-CAM provides class-discriminative heatmaps by utilizing

gradients from convolutional layers and has become a widely adopted technique for visual explanation in CNN-based models [16].

To address these challenges, this study proposes a multimodal deep learning framework that integrates chest radiographic images with demographic metadata, including patient age, sex, and radiographic view position, for multi-label thoracic disease classification. The framework employs a ConvNeXt Large backbone for image feature extraction and a metadata-processing branch for learning complementary patient-specific representations. Image and metadata features are fused prior to classification across fourteen thoracic disease categories, while weighted binary cross-entropy loss is employed to mitigate the impact of severe class imbalance inherent in the dataset. To enhance model transparency, Grad-CAM visualizations are incorporated to facilitate qualitative examination of the image regions contributing to model predictions. Experimental evaluation is performed using patient-level data partitioning and benchmarked against ResNet-152, DenseNet-121, and DenseNet-169. The primary contributions of this study can be summarized as follows:

1. A multimodal deep learning framework is developed to integrate chest radiographs with demographic metadata for large-scale multi-label thoracic disease classification.
2. A ConvNeXt-based architecture combined with weighted binary cross-entropy optimization is proposed to address class imbalance and improve classification performance on the NIH ChestX-ray14 dataset.
3. The framework incorporates Grad-CAM-based visual interpretability and is comprehensively evaluated against widely used deep learning baselines, including ResNet-152, DenseNet-121, and DenseNet-169.

The remainder of this paper is organized as follows. Section 2 reviews related studies on deep learning-based chest X-ray analysis. Section 3 describes the dataset and its main challenges. Section 4 presents the proposed multimodal framework and training configuration. Section 5 reports the experimental results, comparative evaluation, ablation study, interpretability analysis, and discussion. Section 6 concludes the paper and outlines future research directions.

2. Related Work

The rapid advancement of deep learning and the increasing availability of large-scale chest radiograph datasets have significantly accelerated research in automated thoracic disease analysis. Publicly available datasets have played a crucial role in enabling the development and evaluation of deep learning models for disease classification, localization, and clinical decision support. Among the most influential datasets, ChestX-ray8 and its extended version ChestX-ray14 introduced more than 100,000 frontal chest radiographs annotated with fourteen thoracic disease categories, establishing one of the most widely used benchmarks for multi-label chest X-ray classification and weakly supervised localization [7]. Subsequently, CheXpert provided a large-scale dataset with uncertainty-aware labels extracted from radiology reports, enabling more robust model development under label ambiguity [9]. MIMIC-CXR further expanded the field by releasing a large collection of chest radiographs paired with corresponding free-text reports, facilitating multimodal and clinically oriented research [10]. More recently, VinDr-CXR addressed limitations associated with weak labeling by providing expert radiologist annotations and localization information for thoracic abnormalities [11]. Building upon these datasets, numerous deep learning approaches have been proposed for automated thoracic disease detection. One of the most influential studies was CheXNet, introduced by Rajpurkar et al. [8], which

employed DenseNet-121 for pneumonia detection and demonstrated performance comparable to practicing radiologists. Following this work, DenseNet and ResNet architectures became widely adopted baselines for chest radiograph analysis because of their strong feature extraction capabilities and successful transfer learning performance [12][13]. Subsequent studies investigated attention-based mechanisms and feature fusion strategies to improve the recognition of subtle thoracic abnormalities and enhance localization performance [17].

Recent advances in computer vision have introduced more powerful architectures capable of learning richer visual representations. ConvNeXt, proposed by Liu et al. [14], modernized conventional convolutional neural networks by incorporating several design principles inspired by transformer architectures while maintaining the efficiency and scalability of convolutional processing. In parallel, transformer-based models such as Vision Transformer (ViT) [18] and Swin Transformer [19] demonstrated strong performance across a variety of image analysis tasks by modeling long-range spatial dependencies. Nevertheless, the computational efficiency, scalability, and strong representational capacity of ConvNeXt continue to make it an attractive alternative for large-scale medical image analysis. Beyond image-only analysis, increasing attention has been directed toward multimodal learning approaches that combine imaging data with complementary clinical information. Patient demographics, acquisition parameters, and clinical metadata may provide valuable contextual information that is not directly observable from image appearance alone. Several studies have reported improvements in diagnostic performance through the incorporation of metadata into deep learning frameworks [15][20]. More recently, vision-language learning has emerged as a promising research direction in medical imaging. CheXzero [21] demonstrated the feasibility of zero-shot chest X-ray classification using natural language supervision, while MedCLIP [22] introduced contrastive learning between medical images and textual descriptions. Similarly, BioViL [23] explored large-scale vision-language representation learning for biomedical imaging applications. These studies highlight the growing importance of leveraging complementary information sources to improve model generalization and diagnostic performance.

Interpretability has also become an essential requirement for clinical deployment. Although deep learning models can achieve high predictive accuracy, healthcare professionals require transparent explanations that support diagnostic decision-making. Grad-CAM, introduced by Selvaraju et al. [16], remains one of the most widely adopted explainability techniques for visualizing image regions that contribute to model predictions. Numerous chest radiography studies have subsequently employed Grad-CAM and related attention visualization approaches to improve model transparency and facilitate clinical interpretation of automated predictions [24][25]. Table 1 summarizes representative studies in chest radiograph analysis and highlights their methodological characteristics, including metadata utilization, and model interpretability.

Table 1. Comparative analysis of representative deep learning studies for chest X-ray analysis

Study	Year	Dataset	Multimodal Information	Explainability	Main Limitation
CheXNet [8]	2017	ChestX-ray14	No	No	Image-only framework
CheXpert [9]	2019	CheXpert	No	Limited	No demographic information
MIMIC-CXR-based studies [10]	2019–2024	MIMIC-CXR	Partial	Limited	Dataset-specific implementations
VinDr-CXR [11]	2022	VinDr-CXR	No	No	Focused primarily on annotation and localization
ConvNeXt [14]	2022	ImageNet / general vision benchmarks	No	No	Not designed specifically for multimodal chest X-ray diagnosis

Study	Year	Dataset	Multimodal Information	Explainability	Main Limitation
CheXzero [21]	2022	Multiple datasets	No	Limited	Focus on vision-language zero-shot learning
MedCLIP [22]	2022	Multiple datasets	No	Limited	Focus on image-text representation learning
BioViL [23]	2023	MIMIC-CXR	No	Limited	Primarily designed for vision-language pre-training
Proposed Framework	—	ChestX-ray14	Yes (Demographics)	Yes (Grad-CAM)	—

As shown in Table 1, most existing studies focus primarily on image-based analysis and do not incorporate demographic metadata into the learning process. Although explainability techniques have been adopted in some studies, they are often treated as auxiliary components rather than being integrated into multimodal diagnostic frameworks. Consequently, the potential value of combining image information, demographic attributes, and model interpretability within a unified learning framework remains insufficiently explored.

Despite the substantial progress achieved in automated chest radiograph analysis, several challenges remain unresolved. Many existing studies continue to rely primarily on image information while underutilizing demographic attributes that are routinely available in clinical datasets. In addition, severe class imbalance remains a persistent challenge that adversely affects the recognition of rare thoracic abnormalities. Although recent multimodal and vision-language approaches have demonstrated the value of integrating complementary information sources, limited attention has been given to the incorporation of demographic metadata within modern ConvNeXt-based architectures for large-scale multi-label thoracic disease classification. Furthermore, model interpretability is frequently addressed as a separate post hoc analysis rather than being considered alongside multimodal learning design. To address these limitations, the present study introduces a multimodal ConvNeXt-based framework that integrates chest radiographs with demographic metadata for thoracic disease classification while providing interpretable visual explanations through Grad-CAM analysis.

3. Dataset Description

This study utilizes the NIH ChestX-ray14 dataset, one of the most widely adopted public benchmarks for large-scale thoracic disease analysis. Originally introduced by Wang et al. [7], the dataset provides a large-scale collection of chest radiographs together with patient demographic information, making it particularly suitable for investigating multimodal learning strategies that combine image-based and metadata-based representations.

The NIH ChestX-ray14 dataset contains 112,120 frontal chest radiographs acquired from 30,805 unique patients at the National Institutes of Health (NIH) Clinical Center [7]. Images are provided as grayscale PNG files with an original spatial resolution of 1024×1024 pixels. Diagnostic labels were automatically extracted from associated radiology reports using natural language processing techniques, enabling large-scale annotation while resulting in weak image-level labels rather than manually verified expert annotations. The dataset supports multi-label classification across fourteen thoracic disease categories and includes a "No Finding" class representing radiographs without any annotated abnormality [7].

In addition to disease labels, the dataset provides patient demographic information including age, sex, and radiographic view position (AP or PA), enabling investigation of multimodal deep learning approaches that combine image appearance with complementary patient metadata [7]. These metadata attributes are routinely collected during clinical imaging procedures and may contribute valuable contextual information for disease prediction. Their availability makes the NIH ChestX-ray14 dataset

particularly suitable for evaluating metadata-aware deep learning models while facilitating direct comparison with numerous previously published studies using the same benchmark dataset [7][8][12][13].

Table 2 summarizes the main characteristics of the NIH ChestX-ray14 dataset used throughout this study.

Table 2. Characteristics of the NIH ChestX-ray14 dataset used in this study

Attribute	Value
Dataset	NIH ChestX-ray14
Total Images	112,120
Unique Patients	30,805
Disease Categories	14
Additional Class	No Finding
Imaging Modality	Frontal Chest Radiography
Image Format	PNG
Original Resolution	1024 × 1024 pixels
Metadata Available	Age, Sex, View Position
Label Type	NLP-derived Image-Level Labels
Source Institution	NIH Clinical Center

Figure 1 presents representative examples from the NIH ChestX-ray14 dataset, including normal radiographs, single-pathology cases, and images containing multiple coexisting abnormalities. The examples highlight the variability of radiographic appearances across normal cases, single-pathology findings, and images containing multiple abnormalities. While certain abnormalities exhibit distinctive radiographic manifestations, others may appear subtle or overlap with additional pathological findings, making automated interpretation particularly challenging.

Sample Chest X-ray Images from the NIH Dataset

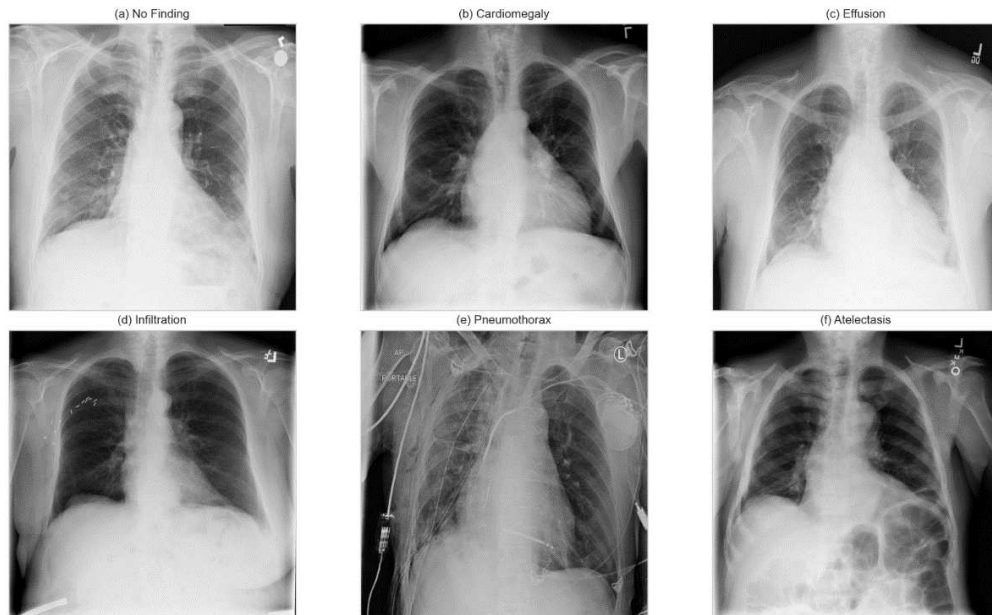


Figure 1. Representative chest X-ray images from the NIH ChestX-ray14 dataset. Top row: normal cases; middle row: single-pathology findings; bottom row: multi-pathology instances [7]

The distribution of demographic and acquisition-related metadata is presented in Figure 2. The patient population spans a broad age range, with a median age of 49 years, reflecting the demographic diversity of the dataset. Male patients account for 56.5% of the radiographs (63,340 images), whereas female

patients represent 43.5% (48,780 images). Regarding image acquisition, posterior–anterior (PA) views constitute 60.0% of the dataset (67,310 images), while anterior–posterior (AP) views account for the remaining 40.0% (44,810 images). These metadata attributes provide clinically relevant contextual information that may complement image-derived representations and improve disease discrimination. Accordingly, age, sex, and view position were incorporated into the proposed multimodal framework as complementary inputs for thoracic disease classification.

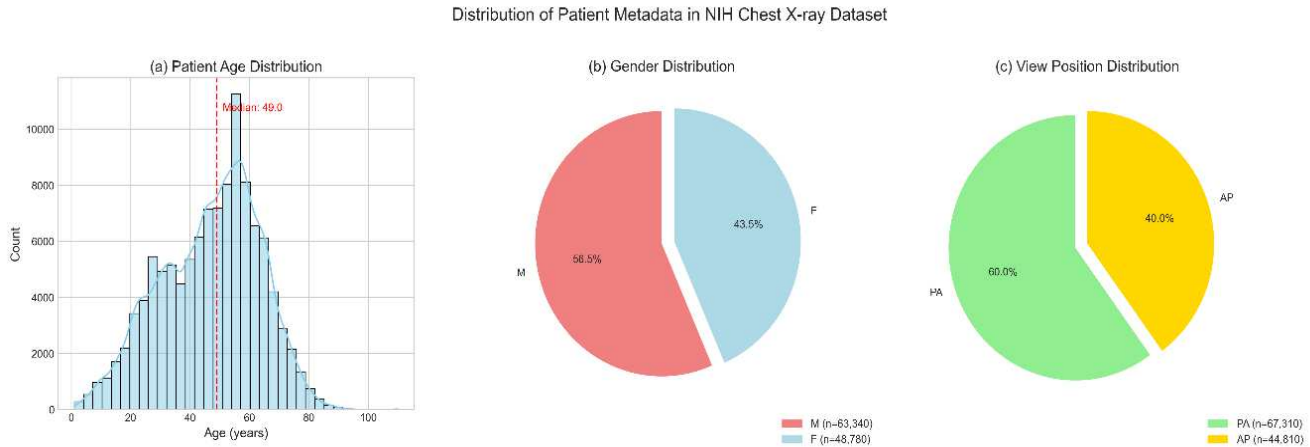


Figure 2. Distribution of patient metadata attributes in the NIH ChestX-ray14 dataset: (a) patient age distribution, (b) sex distribution, and (c) X-ray view position (AP vs. PA) [7]

Figure 3 illustrates the frequency distribution of the fourteen thoracic disease categories in the NIH ChestX-ray14 dataset. A pronounced class imbalance is evident across the dataset. Among pathological findings, Infiltration is the most prevalent category with 19,894 labeled instances, followed by Effusion (13,317) and Atelectasis (11,559). In contrast, Hernia represents the rarest abnormality with only 227 samples. The substantial disparity between common and rare disease categories can bias model learning toward majority classes and reduce sensitivity for underrepresented conditions. This imbalance motivated the adoption of class-weighted optimization during training to improve learning for underrepresented disease categories.

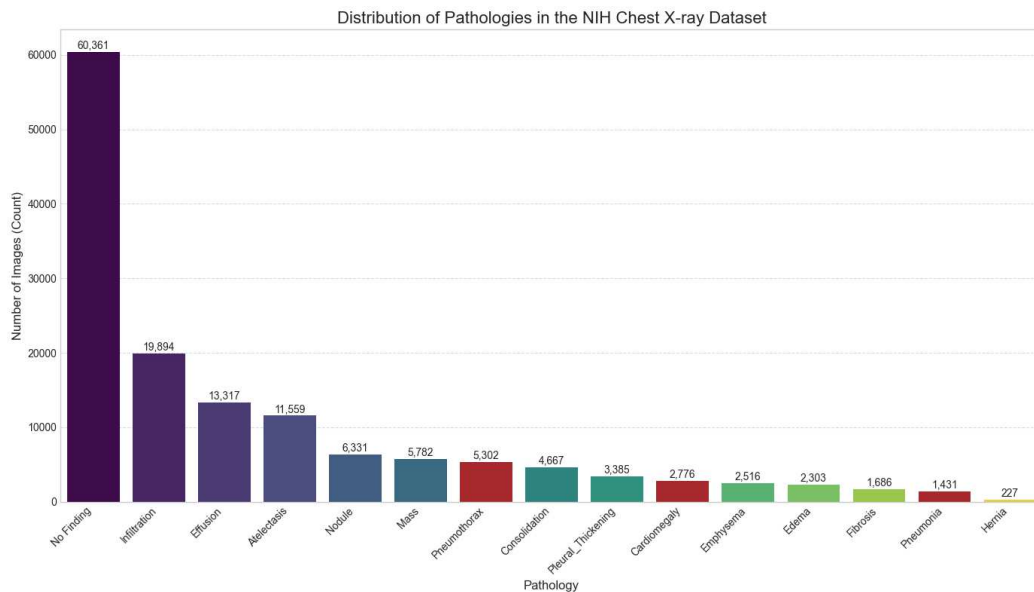


Figure 3. Frequency distribution of the 14 thoracic disease categories in the NIH ChestX-ray14 dataset. A pronounced class imbalance is observed, with Infiltration representing the most prevalent pathological finding and Hernia the rarest abnormality [7]

Despite its scale and clinical relevance, the NIH ChestX-ray14 dataset presents several challenges for automated thoracic disease classification. First, disease labels were automatically extracted from radiology reports using natural language processing techniques, resulting in weak image-level annotations rather than manually verified expert labels [7]. Second, the dataset exhibits severe class imbalance, as illustrated in Figure 3, with common abnormalities represented by substantially more samples than rare pathologies. Third, multiple abnormalities may coexist within a single radiograph, increasing the complexity of multi-label prediction. Finally, because multiple images may originate from the same patient, patient-level data partitioning is necessary to prevent information leakage between training and evaluation subsets. These challenges directly informed the design of the proposed framework. Demographic metadata were incorporated to provide complementary contextual information, weighted binary cross-entropy loss was employed to address class imbalance, and patient-level splitting was enforced throughout the experimental pipeline to ensure reliable performance evaluation.

4. Proposed System

The proposed framework integrates chest radiographic images with patient demographic metadata to improve multi-label thoracic disease classification while providing clinically meaningful visual explanations. The framework combines information derived from radiographs and patient-specific attributes, including age, sex, and radiographic view position, within a unified multimodal learning architecture.

As illustrated in Figure 4, the framework consists of five major stages: (i) image preprocessing and augmentation, (ii) demographic metadata processing, (iii) multimodal feature extraction and fusion, (iv) multi-label disease classification, and (v) Grad-CAM-based visual interpretability. The image processing branch utilizes a ConvNeXt Large backbone pretrained on ImageNet-22K [14] to extract high-level radiographic representations, whereas a dedicated metadata branch learns compact demographic feature embeddings. The outputs of both branches are fused into a joint representation that is subsequently used for multi-label prediction of fourteen thoracic disease categories.

To address the severe class imbalance observed in the NIH ChestX-ray14 dataset, weighted binary cross-entropy optimization is employed during training [29]. Furthermore, strict patient-level data partitioning is enforced throughout the experimental protocol to prevent information leakage and ensure reliable performance estimation. The resulting framework combines strong discriminative capability with enhanced transparency, enabling both accurate disease prediction and clinically interpretable visual explanations.

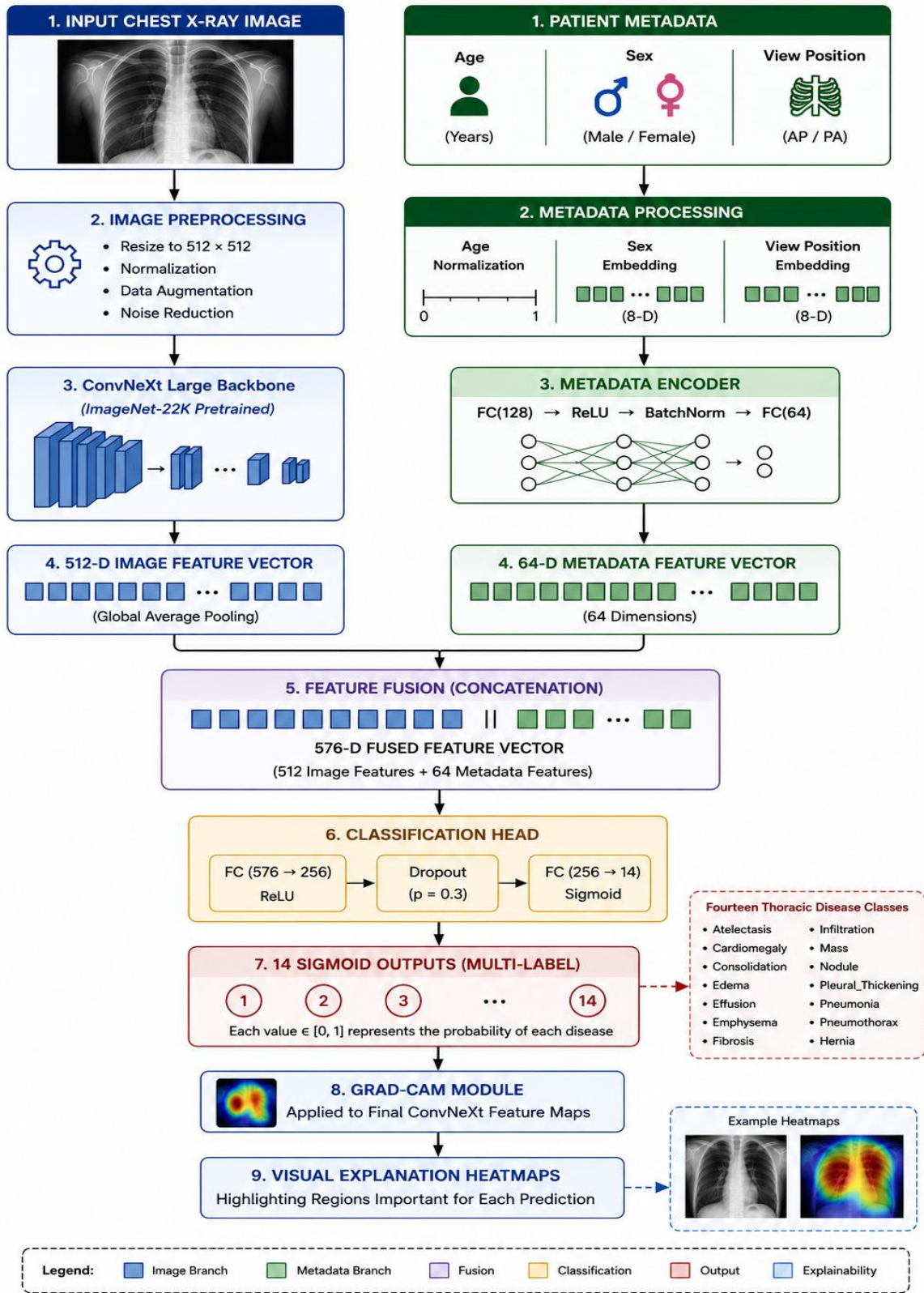


Figure 4. Overview of the proposed multimodal framework for thoracic disease classification. Chest radiographs are processed through a ConvNeXt Large image branch, while demographic metadata (age, sex, and view position) are

processed through a dedicated metadata branch. The extracted features are fused into a joint representation for multi-label disease classification, and Grad-CAM is subsequently applied to generate visual explanations of model predictions.

4.1 Image Preprocessing

Chest radiographs were preprocessed to ensure compatibility with the ConvNeXt Large backbone while preserving diagnostically relevant anatomical information [14]. Given the variability in image acquisition conditions within the NIH ChestX-ray14 dataset, a unified preprocessing pipeline was employed to standardize image characteristics and improve model generalization.

All radiographs were resized to 512×512 pixels, providing a balance between computational efficiency and the preservation of clinically relevant visual details. Pixel intensities were subsequently normalized using the mean and standard deviation statistics associated with ImageNet pretraining (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225] in RGB channel order) to align the input distribution with the pretrained ConvNeXt initialization and facilitate effective transfer learning [14].

To improve robustness against acquisition variability and reduce overfitting, data augmentation was applied exclusively to the training subset. The augmentation strategy incorporated geometric and photometric transformations that mimic realistic variations encountered during clinical image acquisition while preserving pathological patterns and anatomical structures relevant to diagnosis [26]. Representative examples of the applied augmentation techniques are illustrated in Figure 5, while the complete preprocessing and augmentation settings are summarized in Table 3.

Examples of Data Augmentation Techniques

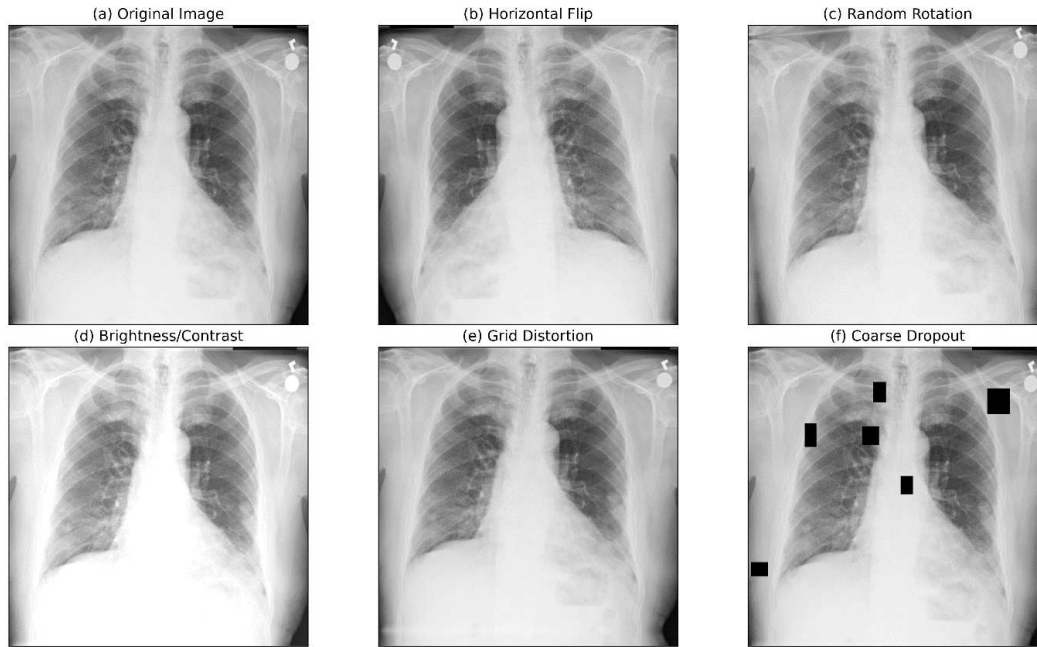


Figure 5. Examples of data augmentation techniques applied to training images. From left to right: original image, horizontal flip, rotation, brightness adjustment, and Gaussian noise addition.

Table 3. Image preprocessing and augmentation settings employed during model training

Parameter	Value
Input resolution	512×512
ImageNet mean (RGB)	(0.406, 0.456, 0.485)
ImageNet std (RGB)	(0.225, 0.224, 0.229)
Horizontal flip probability	0.5

Parameter	Value
Rotation range	$\pm 15^\circ$
Translation range	$\leq 5\%$
Brightness adjustment	$\pm 20\%$
Contrast scaling	0.8–1.2
Gaussian noise standard deviation	≤ 0.01
Random resized crop scale	0.85–1.0
Application stage	Training set only

4.2 Metadata Preprocessing

In addition to radiographic images, the proposed framework incorporates patient-specific metadata to provide complementary contextual information for thoracic disease classification. Three metadata attributes available within the NIH ChestX-ray14 dataset were utilized: patient age, patient sex, and radiographic view position (AP or PA) [7]. These variables were selected because they are routinely available in clinical practice and may influence disease prevalence, radiographic appearance, and image acquisition characteristics.

Patient age was treated as a continuous numerical variable and normalized to the range (0,1) prior to model training. Normalization was applied to ensure numerical consistency with the remaining input features and to facilitate stable optimization during network training.

Patient sex and radiographic view position were treated as categorical variables. After categorical encoding, each category was mapped to a learnable embedding representation. Unlike one-hot encoding, embedding layers allow the network to learn compact task-specific feature representations while preserving relationships among categorical attributes [27].

The processed age feature and categorical embeddings were concatenated to form a unified metadata vector. This vector was subsequently passed through a lightweight fully connected network consisting of two dense layers with ReLU activation and batch normalization, producing a compact 64-dimensional metadata representation [28]. The resulting embedding captures demographic and acquisition-related information that complements the image-based features extracted by the ConvNeXt backbone and serves as an input to the multimodal fusion stage described in the following section.

4.3 Multimodal Feature Fusion

The dataset analysis presented in Section 3 demonstrated that patient age, sex, and radiographic view position provide clinically relevant contextual information that may complement image-based representations. Motivated by these observations, the proposed framework adopts a multimodal architecture that jointly exploits radiographic and demographic information to improve thoracic disease classification performance. As illustrated in Figure 4, the image branch employs a ConvNeXt Large backbone pretrained on ImageNet-22K to extract high-level visual representations from chest radiographs [14]. Following global average pooling and dropout regularization ($p = 0.3$), the extracted visual features are projected into a 512-dimensional image embedding.

In parallel, the metadata branch processes patient age together with the categorical attributes corresponding to patient sex and radiographic view position. After normalization and embedding, these features are passed through two fully connected layers with ReLU activation and batch normalization, generating a compact 64-dimensional metadata representation. The image and metadata representations are subsequently integrated through feature-level fusion. The multimodal fusion operation is defined as:

$$F_{Fusion} = \text{Concat}(F_I, F_M) \quad (1)$$

where F_I and F_M represent the image and metadata feature embeddings, respectively. The resulting fused representation contains 576 features and serves as the input to the subsequent classification module.

As defined in Equation (1), feature-level fusion preserves information from both modalities while enabling the network to learn interactions between radiographic patterns and patient-specific contextual attributes within a shared latent space. Compared with late-fusion strategies, this approach allows the classifier to exploit correlations between visual findings and demographic factors during feature learning. By integrating image-derived and metadata-derived representations, the proposed framework captures complementary diagnostic cues that may not be available from either modality independently. The resulting multimodal representation is subsequently forwarded to the classification module for multi-label thoracic disease prediction.

4.4 Multi-Label Classification and Optimization

The fused multimodal representation obtained from the image and metadata branches is subsequently forwarded to the classification module for thoracic disease prediction. Since multiple abnormalities may coexist within a single chest radiograph, the problem is formulated as a multi-label classification task rather than a mutually exclusive multi-class classification problem [29]. The classification module consists of two fully connected layers. The first layer transforms the 576-dimensional fused representation into a 256-dimensional latent representation using a Rectified Linear Unit (ReLU) activation function. The resulting features are then mapped to fourteen independent output neurons corresponding to the thoracic disease categories included in the NIH ChestX-ray14 dataset. A sigmoid activation function is applied to each output node to estimate the probability of individual pathologies independently. The prediction process is defined as:

$$P = \sigma(WF_{\text{Fusion}} + b) \quad (2)$$

where F_{Fusion} denotes the fused multimodal representation obtained from Equation (1), W and b represent the trainable weight matrix and bias vector, respectively, $\sigma(\cdot)$ denotes the sigmoid activation function, and P corresponds to the vector of predicted disease probabilities. As shown in Figure 3, the NIH ChestX-ray14 dataset exhibits substantial class imbalance, with common abnormalities appearing considerably more frequently than rare pathological findings. To mitigate the influence of this imbalance during training, a weighted binary cross-entropy (WBCE) loss function was adopted. Class-specific weights were computed using the inverse positive-sample ratio within the training subset, thereby increasing the contribution of underrepresented disease categories during optimization [29].

The optimization objective is defined as:

$$L_{WBCE} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^C w_c [y_{ic} \log(\hat{p}_{ic}) + (1 - y_{ic}) \log(1 - \hat{p}_{ic})] \quad (3)$$

where N denotes the total number of training samples, C represents the number of disease classes, y_{ic} is the ground-truth label, $\hat{p}_{ic}$ is the predicted probability for class c , and w_c corresponds to the class-specific weighting factor. By assigning larger weights to minority classes, the loss function encourages the model to learn discriminative representations for rare abnormalities while maintaining strong performance on more prevalent diseases.

All models were initialized using ImageNet-22K pretrained weights and subsequently fine-tuned end-to-end on the NIH ChestX-ray14 dataset. To ensure a fair comparison, all baseline architectures (ResNet-152, DenseNet-121, and DenseNet-169) were trained under conditions identical to the proposed framework, including the same input resolution (512×512 pixels), data augmentation pipeline, weighted binary cross-entropy loss function, optimizer configuration, and patient-level data partitioning. Patient-level stratified splitting was enforced throughout the experimental protocol to prevent information leakage between training and evaluation subsets. The dataset was divided into 70% training, 15% validation, and 15% testing partitions. Optimization was performed using the Adam optimizer [30] with adaptive learning-rate scheduling and early stopping based on validation AUC. The complete training configuration and optimization settings are summarized in Table 4.

Table 4. Training configuration and optimization settings of the proposed framework

Hyperparameter / Setting	Value
Framework	PyTorch 2.0
GPU	NVIDIA A100 (40 GB VRAM)
Input resolution	512 × 512 pixels
Optimizer	Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$)
Initial learning rate	1×10^{-4}
LR scheduler	ReduceLROnPlateau (factor = 0.5, patience = 5 epochs)
Batch size	32
Training epochs	50 (early stopping: patience = 10, monitor: val. AUC)
Loss function	Weighted binary cross-entropy (class weights = inverse positive-sample ratio)
Dropout rate	0.3 (image branch FC layer)
Pre-trained weights	ImageNet-22K
Data split	70% train / 15% val / 15% test (patient-level stratified)
Weight decay	1×10^{-5}
Random seed	42

4.5 Grad-CAM-Based Visual Interpretability

To improve the interpretability of the proposed framework and facilitate clinical understanding of model predictions, Gradient-weighted Class Activation Mapping (Grad-CAM) was employed as a post-hoc visual explanation technique. Grad-CAM generates class-specific activation maps by utilizing the gradients of the target class with respect to the final convolutional feature maps of the ConvNeXt backbone. The resulting heatmaps highlight image regions that contribute most strongly to the model’s decision for each predicted pathology [16]. For each predicted thoracic abnormality, the generated activation map is projected onto the corresponding chest radiograph to identify diagnostically relevant regions. Areas with higher activation values indicate image regions that exert greater influence on the predicted disease probability. This enables qualitative assessment of whether the model focuses on clinically meaningful anatomical structures and pathological manifestations. Unlike model-specific attention mechanisms, Grad-CAM provides visual explanations without modifying the underlying classification architecture or affecting model performance. By identifying radiographically relevant regions associated with the predicted pathology, Grad-CAM facilitates qualitative assessment of model behavior and enhances the transparency and clinical interpretability of the proposed framework.

5. Results and Discussion

The proposed framework was evaluated on the held-out test set using three complementary perspectives: multi-label classification performance, comparative analysis against baseline architectures, and Grad-

CAM-based visual interpretability. The test set represented 15% of the patient-level partitioned NIH ChestX-ray14 dataset and included 16,741 images from patients not used during training or validation.

5.1 Classification Performance of the Proposed ConvNeXt Model

The proposed ConvNeXt-based framework demonstrated strong classification performance across the fourteen thoracic disease categories included in the NIH ChestX-ray14 dataset. The receiver operating characteristic (ROC) curves shown in Figure 6 indicate excellent discriminative capability across all classes, yielding a macro-average AUC of 96.88%. As illustrated in Figure 6, the highest AUC values were obtained for Hernia (0.999), Emphysema (0.990), Pneumothorax (0.988), and Cardiomegaly (0.986). High discrimination performance was also observed for Effusion (0.970), Pleural Thickening (0.967), Nodule (0.964), Consolidation (0.959), Atelectasis (0.957), and Pneumonia (0.951). All disease categories achieved AUC values above 0.90, indicating consistently strong discrimination performance across the evaluated pathologies.

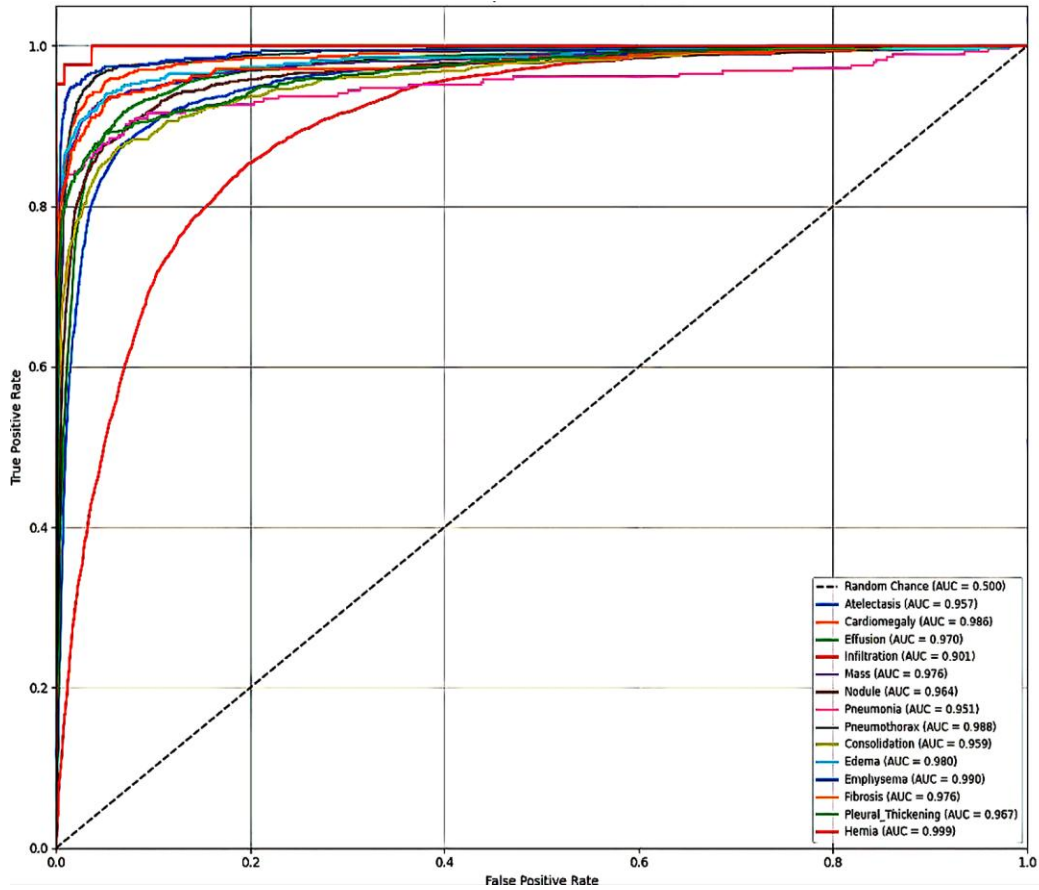


Figure 6. Receiver operating characteristic (ROC) curves of the proposed multimodal framework for the fourteen thoracic disease categories on the held-out test set.

The class-wise performance metrics presented in Figure 7 provide additional insight into model behavior. The highest F1-scores were achieved for Hernia (0.95), Emphysema (0.92), Pneumonia (0.91), Edema (0.91), and Effusion (0.91). Strong performance was also observed for Mass (0.90), Atelectasis (0.89), and Cardiomegaly (0.88). In contrast, Infiltration recorded the lowest F1-score (0.81), which may

be attributed to its heterogeneous radiographic appearance and visual overlap with several pulmonary abnormalities. Sensitivity values ranged from 0.83 to 0.94 across all disease categories, while specificity remained consistently high, ranging from 0.90 to 1.00. Precision values were similarly robust, reaching 0.97 for Pneumonia and 0.95 for Fibrosis. These results indicate that the proposed framework maintains a favorable balance between sensitivity and specificity while minimizing false-positive predictions.

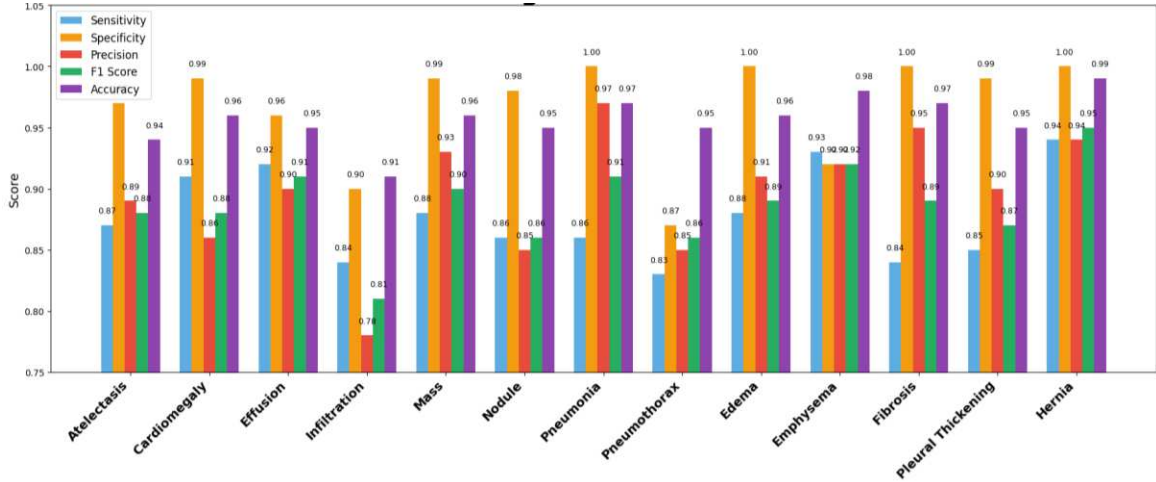


Figure 7. Class-wise sensitivity, specificity, precision, F1-score, and accuracy achieved by the proposed framework on the held-out test set

The confusion matrices shown in Figure 8 further support the quantitative findings. Most disease categories exhibited strong separation between positive and negative cases, with relatively limited misclassification rates. Higher classification accuracy was observed for categories such as Pneumothorax, Cardiomegaly, Effusion, and Mass, whereas greater confusion was observed for Infiltration and Nodule, reflecting the increased complexity associated with subtle or overlapping radiographic patterns. Collectively, the results presented in Figures 6–8 demonstrate that the proposed multimodal ConvNeXt framework achieves accurate and reliable multi-label classification across a diverse range of thoracic abnormalities.

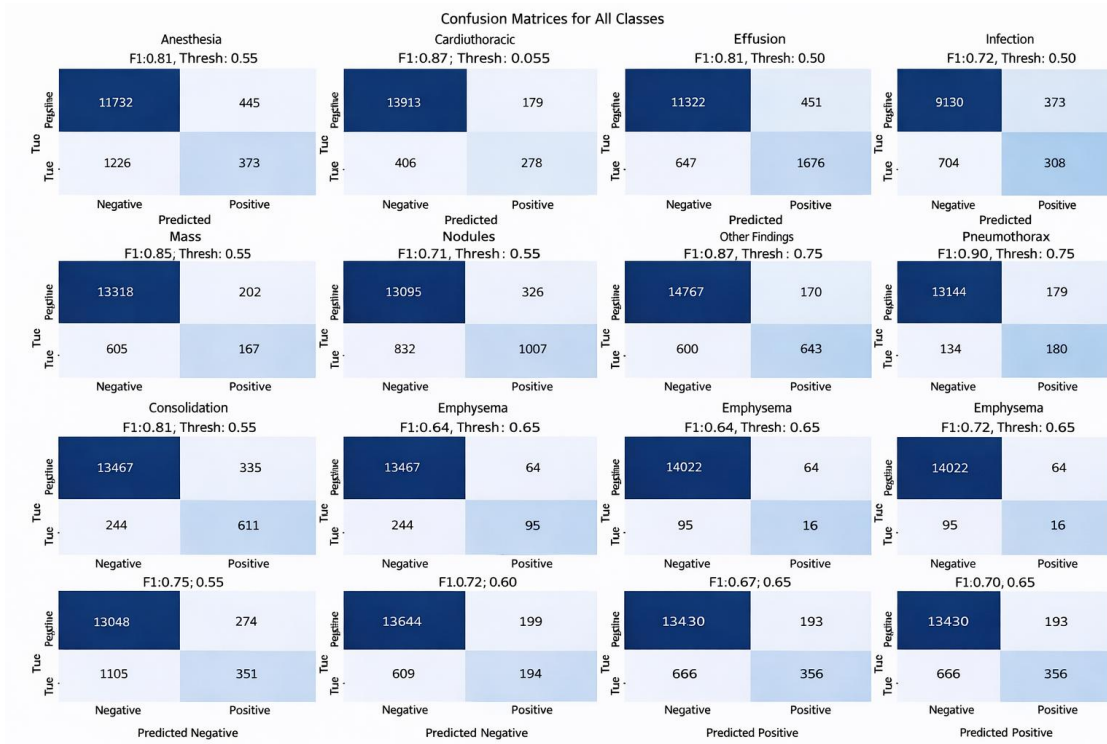


Figure 8. Class-wise confusion matrices of the proposed framework across the fourteen thoracic disease categories on the held-out test set

5.2 Comparative Performance Analysis

To assess the effectiveness of the proposed multimodal framework, its performance was compared against three widely used deep learning architectures: ResNet-152, DenseNet-121, and DenseNet-169. The corresponding ROC curves for the baseline models are presented in Figures 9–11, while the overall comparison is summarized in Figure 12 and Table 5.

The proposed ConvNeXt framework achieved the highest overall performance, obtaining a macro-average AUC of 96.88%. In comparison, DenseNet-121, DenseNet-169, and ResNet-152 achieved macro-average AUC values of 87.76%, 87.21%, and 85.02%, respectively. These results demonstrate a consistent improvement across all evaluated architectures and confirm the effectiveness of the proposed multimodal framework.

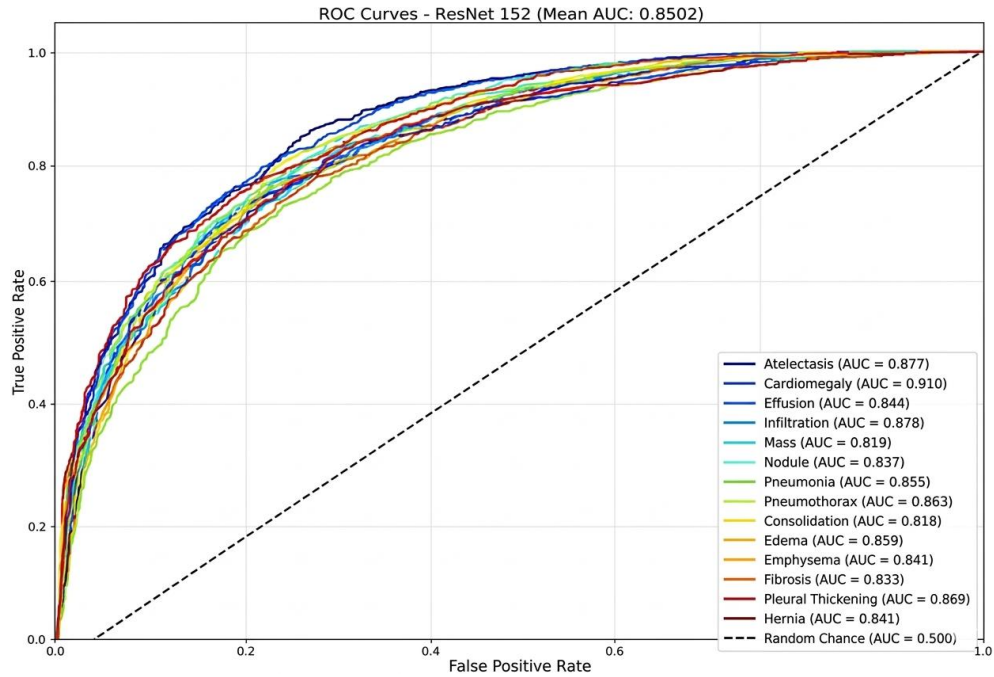


Figure 9. Receiver operating characteristic (ROC) curves for the fourteen thoracic disease categories predicted by the ResNet-152 model on the held-out test set

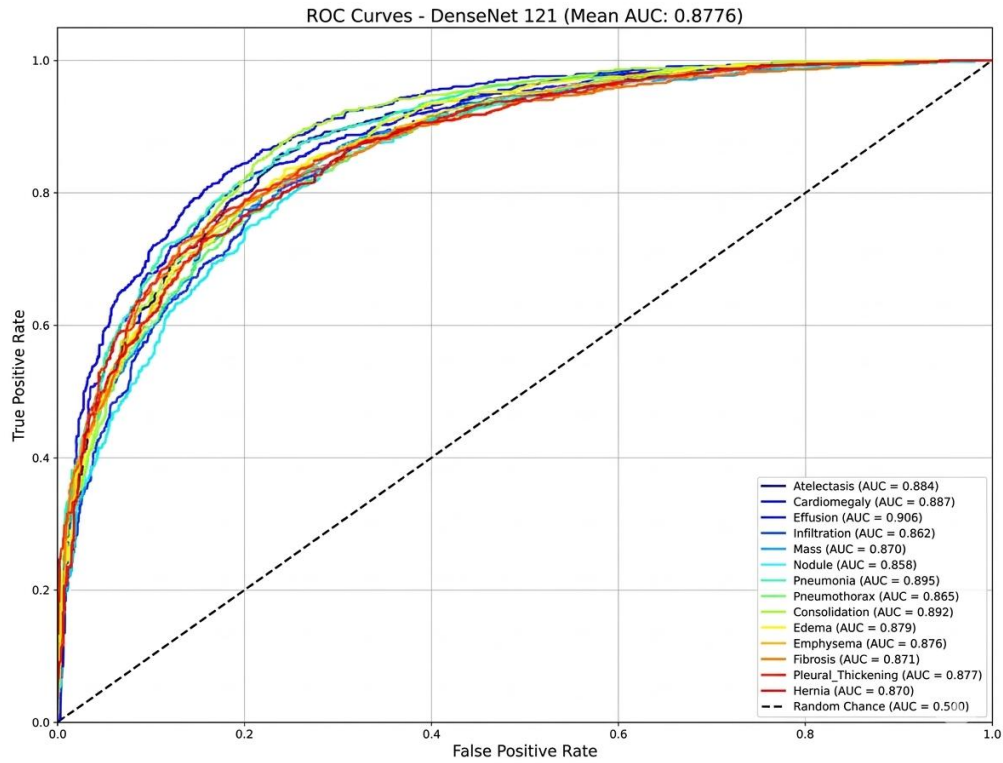


Figure 10. Receiver operating characteristic (ROC) curves for the fourteen thoracic disease categories predicted by the DenseNet-121 model on the held-out test set

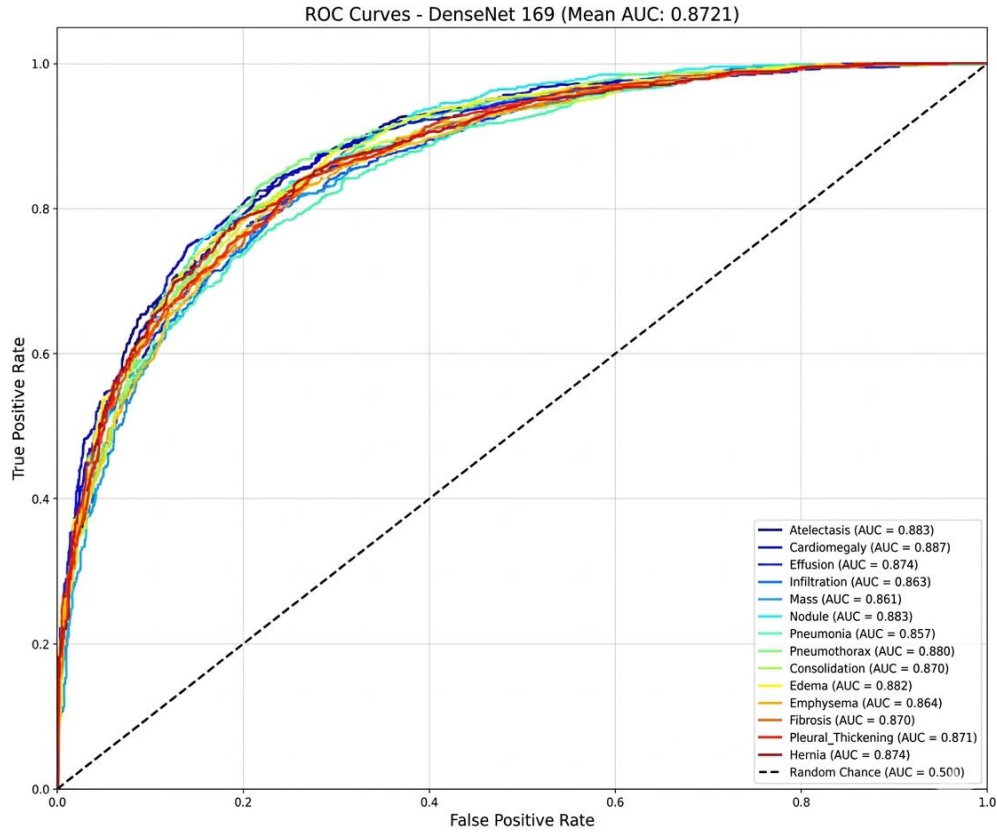


Figure 11. Receiver operating characteristic (ROC) curves for the fourteen thoracic disease categories predicted by the DenseNet-169 model on the held-out test set

Figure 12 provides a direct comparison of the macro-average AUC scores achieved by the four evaluated architectures. The proposed framework outperformed DenseNet-121 by 9.12 percentage points, DenseNet-169 by 9.67 percentage points, and ResNet-152 by 11.86 percentage points.

The quantitative results summarized in Table 5 further confirm the superiority of the proposed framework across multiple evaluation metrics. In addition to achieving the highest AUC, the model also obtained the highest mean F1-score (0.84), mean sensitivity (0.83), and mean specificity (0.94). These findings indicate that the proposed multimodal architecture not only improves discrimination capability but also maintains a favorable balance between true-positive and true-negative predictions.

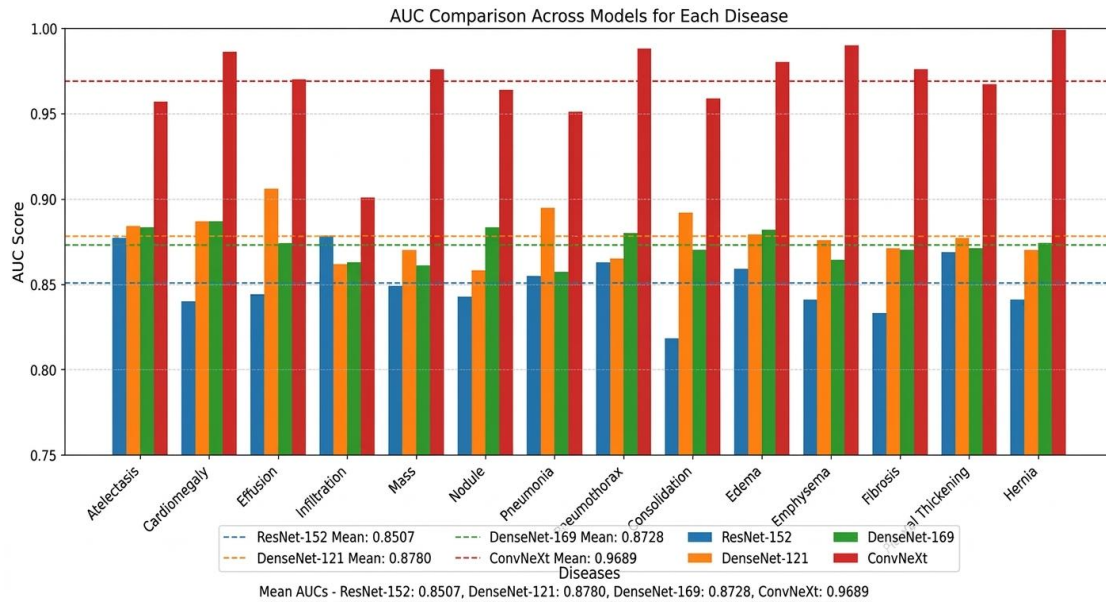


Figure 12. Comparison of the macro-average AUC scores achieved by the four evaluated architectures on the held-out test set

Table 5. Comparative classification performance of the evaluated architectures on the held-out test set

Model	Mean AUC (%)	Mean F1	Mean Sensitivity	Mean Specificity	Parameters (M)
ConvNeXt Large + Demographics (Proposed)	96.88	0.84	0.83	0.94	197.8
DenseNet-121	87.76	0.74	0.73	0.88	7.0
DenseNet-169	87.21	0.73	0.72	0.88	14.1
ResNet-152	85.02	0.71	0.70	0.87	60.2

The observed performance improvements can be attributed to three complementary factors. First, ConvNeXt Large provides a stronger feature extraction capability than the evaluated baseline architectures. Second, demographic metadata introduce complementary patient-specific information that is unavailable to image-only models. Third, weighted binary cross-entropy reduces the adverse effect of severe class imbalance by increasing the contribution of underrepresented disease categories during optimization.

Although the proposed framework contains substantially more trainable parameters than the baseline models, the observed improvements are unlikely to be explained solely by model capacity. Performance gains were consistently observed across all evaluation metrics and disease categories, suggesting that multimodal feature fusion and imbalance-aware optimization contribute meaningfully to the overall classification performance. Table 6 presents a complete per-class AUC breakdown for all fourteen thoracic disease categories across the four evaluated architectures, confirming that the proposed framework achieves superior discrimination for every individual pathology.

Table 6. Per-class AUC scores of the proposed framework and baseline architectures on the held-out test set

Disease / Pathology	ConvNeXt Large + Demo. (Proposed)	ResNet-152	DenseNet-121	DenseNet-169
Atelectasis	0.957	0.877	0.884	0.883
Cardiomegaly	0.986	0.910	0.887	0.887
Consolidation	0.959	0.818	0.892	0.870

Edema	0.980	0.859	0.879	0.882
Effusion	0.970	0.844	0.906	0.874
Emphysema	0.990	0.841	0.876	0.864
Fibrosis	0.976	0.833	0.871	0.870
Hernia	0.999	0.841	0.870	0.874
Infiltration	0.901	0.878	0.862	0.863
Mass	0.976	0.819	0.870	0.861
Nodule	0.964	0.837	0.858	0.883
Pleural Thickening	0.967	0.869	0.877	0.871
Pneumonia	0.951	0.855	0.895	0.857
Pneumothorax	0.988	0.863	0.865	0.880
Mean AUC (%)	96.88	85.02	87.76	87.21

5.3 Ablation Study and Model Interpretability

An ablation study was conducted to examine the contribution of the main components of the proposed framework. Weighted binary cross-entropy optimization and demographic metadata integration were progressively incorporated into the ConvNeXt Large baseline. The resulting changes in macro-average AUC are summarized in Table 7. The baseline ConvNeXt Large model achieved a macro-average AUC of 94.73%. Incorporating weighted binary cross-entropy increased the AUC to 95.81%, corresponding to an improvement of 1.08 percentage points. This improvement indicates that imbalance-aware optimization helped reduce the effect of the skewed class distribution in the NIH ChestX-ray14 dataset.

The integration of demographic metadata further increased the macro-average AUC to 96.12%, suggesting that patient-specific contextual information contributes additional diagnostic value beyond image-derived features alone. The complete framework achieved the highest performance, reaching a macro-average AUC of 96.88%. The progressive improvement across configurations indicates that both metadata fusion and weighted binary cross-entropy contributed to the final classification performance.

Table 6. Ablation analysis of the proposed framework

Configuration	Mean AUC (%)
ConvNeXt Large only	94.73
ConvNeXt + Weighted BCE	95.81
ConvNeXt + Demographic Metadata	96.12
Proposed Full Framework	96.88

To further examine model behavior, Grad-CAM heatmaps were generated for representative correctly classified cases from each thoracic disease category. The visualizations revealed that different disease predictions were associated with anatomically plausible regions. For example, Pneumothorax predictions primarily highlighted peripheral lung regions, whereas Cardiomegaly activation maps concentrated around the enlarged cardiac silhouette. In cases of Pleural Effusion, the strongest responses were consistently observed in the dependent lower lung fields, corresponding to the expected location of pleural fluid. Similar localization patterns were observed for the remaining disease categories, indicating that the network extracted pathology-specific imaging features rather than relying on diffuse image responses. Although Grad-CAM provides qualitative rather than quantitative evidence, these visualizations improve understanding of the model's decision process and demonstrate that the learned representations are consistent with established radiographic characteristics.

5.4 Discussion

The experimental results show that combining multimodal feature fusion with imbalance-aware optimization improved thoracic disease classification performance on the NIH ChestX-ray14 dataset. Across the evaluated architectures, the proposed framework achieved the highest macro-average AUC, mean F1-score, sensitivity, and specificity, indicating strong discrimination capability and balanced prediction performance. The comparative analysis and ablation study suggest that the observed gains were not solely due to the ConvNeXt backbone. Demographic metadata integration and weighted binary cross-entropy optimization both contributed to the final performance. These findings support the value of incorporating patient-specific contextual information alongside image-derived features for multi-label chest radiograph classification. Clinically, this suggests that systems integrating routinely collected patient data with radiographic analysis may offer more reliable diagnostic support than image-only approaches, particularly for pathologies whose prevalence or radiographic appearance varies with patient demographics. Disease-level analysis reveals noteworthy patterns: Hernia achieved the highest AUC (0.999), consistent with its distinctive radiographic appearance and the known influence of patient age and sex on hernia prevalence. Emphysema (0.990) and Pneumothorax (0.988) also showed near-perfect discrimination, likely reflecting their characteristic radiographic manifestations combined with the complementary value of view position metadata. In contrast, Infiltration recorded the lowest F1-score (0.81), attributable to its heterogeneous visual appearance and frequent overlap with Consolidation and Edema — conditions where demographic context provides less discriminative power than distinct visual features.

Several limitations should be acknowledged. First, the NIH ChestX-ray14 dataset relies on NLP-derived labels, which may introduce varying levels of label noise across disease categories; this may disproportionately affect rare conditions such as Hernia and Pneumonia, where annotation errors are more consequential. Second, all experiments were conducted using a single dataset, and external validation on independent multi-center cohorts remains necessary to assess model generalizability across different scanner types, acquisition protocols, and patient populations. Third, the relatively large number of trainable parameters in ConvNeXt Large (197.8M) may limit deployment in resource-constrained clinical environments, motivating future investigation of knowledge distillation or pruning strategies. Fourth, the demographic metadata utilized in this study is limited to age, sex, and view position; incorporation of additional clinical attributes such as smoking history, prior diagnoses, or laboratory findings could further improve predictive performance.

Future research should investigate lightweight multimodal architectures such as knowledge distillation from ConvNeXt Large into compact student models, external multi-center validation on independent cohorts, and the incorporation of additional patient metadata sources including clinical history and laboratory findings. Furthermore, exploring uncertainty quantification methods would improve the reliability of automated predictions in clinical settings where model confidence is as important as accuracy.

6. Conclusion

This study presented a multimodal deep learning framework that integrates a ConvNeXt Large backbone with patient demographic metadata — including age, sex, and radiographic view position — for multi-label classification of fourteen thoracic diseases using the NIH ChestX-ray14 dataset. To address the severe class imbalance inherent in the dataset, weighted binary cross-entropy optimization was incorporated during model training, while patient-level data partitioning was employed to ensure reliable performance evaluation. Experimental results demonstrated that the proposed framework achieved a macro-average AUC of 96.88%, outperforming ResNet-152 (85.02%), DenseNet-121 (87.76%), and DenseNet-169 (87.21%) under the same experimental protocol. Ablation analysis further confirmed that

both demographic metadata integration and imbalance-aware optimization contributed to the observed performance improvements. In addition, Grad-CAM visualizations provided qualitative evidence that the model focused on anatomically relevant image regions, supporting the interpretability of the proposed framework. Overall, the results demonstrate that integrating demographic metadata with deep image representations can improve multi-label thoracic disease classification while preserving model transparency through visual explanation. These findings highlight the potential of multimodal learning to support the development of reliable computer-aided diagnostic systems for chest radiography.

Future work will focus on validating the proposed framework using independent external datasets, investigating lightweight multimodal architectures for efficient clinical deployment, incorporating additional patient-specific clinical information, and exploring federated learning strategies to enable privacy-preserving multi-institutional model development.

Acknowledgment

The authors extend their appreciation to Prince Sattam bin Abdulaziz University for funding this research work through the project number (PSAU/2025/1/38421).

DECLARATION

Funding

The authors extend their appreciation to Prince Sattam bin Abdulaziz University for funding this research work through the project number (PSAU/2025/1/38421).

Code Availability

The source code supporting the findings of this study is available from the corresponding author upon reasonable request. A public repository will be made available upon acceptance of this manuscript.

Conflict Of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Consent for publication

Not applicable.

Consent to Participate

Not applicable.

Ethics approval and consent to participate

Not applicable.

Availability of data and materials

The data used in this study are publicly available from the National Institutes of Health (NIH) Clinical Center: doi: 10.1109/CVPR.2017.369. Additional processed data and implementation details are available from the corresponding author upon reasonable request.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Noha E. ElAttar led the conceptualization of the study, designed the methodology, supervised the overall research process, and prepared the original manuscript draft, in addition to conducting critical review and editing. Samia M. Abdelhaleem contributed to data curation, software implementation, experimental investigation, and validation of the results, and participated in manuscript review and editing. Sultan M. Aldossary supported the formal analysis, contributed to methodological refinement, prepared visualizations, and assisted in reviewing and editing the manuscript. Magda M.

Algameel provided domain expertise related to medical and clinical interpretation, contributed to validation and resource support, and participated in manuscript review and editing. All authors have read and approved the final version of the manuscript and agree to be accountable for all aspects of the work.

References

- [1] K. C. Santosh and S. Antani, "Automated chest X-ray screening: Can lung region symmetry help detect pulmonary abnormalities?," *IEEE Transactions on Medical Imaging*, vol. 37, no. 5, pp. 1168–1177, May 2018, doi: 10.1109/TMI.2017.2747503.
- [2] D. S. Kermany, M. Goldbaum, W. Cai, et al., "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131.e9, Feb. 2018, doi: 10.1016/j.cell.2018.02.010.
- [3] A. B. Rimmer, "Radiologist shortage leaves patient care at risk, warns royal college," *BMJ*, vol. 359, p. j4683, 2017, doi: 10.1136/bmj.j4683.
- [4] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, no. 1, pp. 44–56, Jan. 2019, doi: 10.1038/s41591-018-0300-7.
- [5] J. A. Brady, "Error and discrepancy in radiology: Inevitable or avoidable?," *Insights into Imaging*, vol. 8, no. 1, pp. 171–182, Feb. 2017, doi: 10.1007/s13244-016-0534-1.
- [6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
- [7] 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)*, 2017, pp. 2097–2106, doi: 10.1109/CVPR.2017.369.
- [8] P. Rajpurkar, J. Irvin, K. Zhu, et al., "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," *arXiv preprint*, arXiv:1711.05225, 2017.
- [9] J. Irvin, P. Rajpurkar, M. Ko, et al., "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 1, pp. 590–597, 2019, doi: 10.1609/aaai.v33i01.3301590.
- [10] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, et al., "MIMIC-CXR: A de-identified publicly available database of chest radiographs with free-text reports," *Scientific Data*, vol. 6, no. 1, p. 317, 2019, doi: 10.1038/s41597-019-0322-0.
- [11] H. Q. Nguyen, K. Lam, L. T. Le, et al., "VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations," *Scientific Data*, vol. 9, no. 1, p. 429, 2022, doi: 10.1038/s41597-022-01498-w.
- [12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
- [13] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4700–4708, doi: 10.1109/CVPR.2017.243.
- [14] 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 Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11976–11986, doi: 10.1109/CVPR52688.2022.01167.
- [15] T. K. K. Ho and J. Gwak, "Multiple Feature Integration for Classification of Thoracic Disease in Chest Radiography," *Applied Sciences*, vol. 9, no. 19, p. 4130, 2019, doi: 10.3390/app9194130.
- [16] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626, doi: 10.1109/ICCV.2017.74.
- [17] R. Gu et al., "CA-Net: Comprehensive Attention Convolutional Neural Networks for Explainable Medical Image Segmentation," *IEEE Transactions on Medical Imaging*, vol. 40, no. 2, pp. 699–711, 2021.
- [18] A. Dosovitskiy et al., "An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale," *ICLR*, 2021.
- [19] Z. Liu et al., "Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows," *ICCV*, 2021, doi: 10.1109/ICCV48922.2021.00986.
- [20] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE TPAMI*, vol. 41, no. 2, pp. 423–443, 2019.
- [21] E. Tiu et al., "Expert-Level Detection of Pathologies from Unannotated Chest X-ray Images via Self-Supervised Learning," *Nature Biomedical Engineering*, vol. 6, pp. 1399–1406, 2022.
- [22] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "MedCLIP: Contrastive Learning from Unpaired Medical Images and Text," *EMNLP*, 2022.
- [23] B. Boecking et al., "Making the Most of Text Semantics to Improve Biomedical Vision-Language Processing," *ECCV*, 2022.

- [24] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and Explainability of Artificial Intelligence in Medicine," *WIREs Data Mining and Knowledge Discovery*, vol. 9, no. 4, e1312, 2019.
- [25] S. Saporta, A. Gui, A. Agrawal, et al., "Benchmarking saliency methods for chest X-ray interpretation," *Nature Machine Intelligence*, vol. 4, pp. 867–878, 2022, doi: 10.1038/s42256-022-00536-4.
- [26] Shorten, C., and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, p. 60, 2019, doi: 10.1186/s40537-019-0197-0.
- [27] Y. Guo and F. Berkhahn, "Entity Embeddings of Categorical Variables," *arXiv preprint*, arXiv:1604.06737, 2016.
- [28] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," in *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, Lille, France, 2015, pp. 448–456.
- [29] J. C. Dunnmon, D. Yi, C. P. Langlotz, D. Ré, D. L. Rubin, and M. P. Lungren, "Assessment of Convolutional Neural Networks for Multi-Label Classification of Chest Radiographs," *Radiology: Artificial Intelligence*, vol. 1, no. 2, e180013, 2019, doi: **10.1148/ryai.2019180013**.
- [30] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, 2015.