Interpretable Chest X-Ray Diagnosis with Deep Learning

Edrin Hasaj [1] Haris Aljic [2] Abdullah Siddiqui [3] Ibrahim Youssef [4]

[1]University of Toronto, Department of Computer Science

Problem and Dataset Description

The goal of this project is to develop a deep learning system to classify and diagnose thoracic diseases from chest X-ray images.

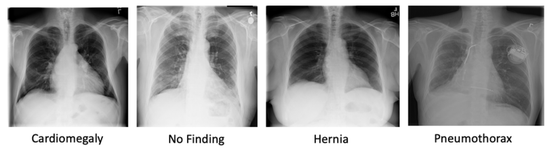


Figure 1. Sample chest X-ray from the NIH ChestX-ray14 dataset.

We use the NIH [5] ChestX-ray14 dataset, a large public medical imaging dataset with **112,120** frontal chest X-rays from **30,805** unique patients and **14** disease labels, including Pneumonia, Atelectasis, Cardiomegaly, Mass, Nodule, Effusion, and Infiltration.

Challenges and Limitations of the Dataset:

- **Class imbalance:** Some diseases dominate the dataset, making rare diseases harder to learn.
- **Multi-label complexity:** Each image may have multiple disease labels, adding to classification difficulty.
- **Noisy data:** Some images contain artifacts or low contrast.

Model Exploration

We employ both Convolutional Neural Networks (CNNs) and CNN-Transformer hybrid models for disease classification on chest X-rays.

1. CNNs: CNNs are effective at capturing local patterns and hierarchical features in medical images. *Models used:* DenseNet121, VGG19

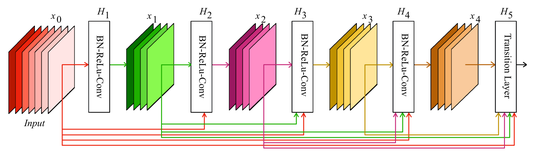


Figure 2. DenseNet121 Architecture [3].

2. CNN-Transformer Hybrids: These models combine the strengths of convolution (local feature extraction) and self-attention (global context modeling) for enhanced performance. *Models used:* MaxViT, CoAtNet, ConvNeXt.

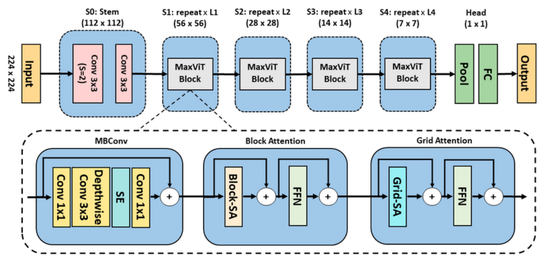


Figure 3. MaxViT Architecture [4].

Dataset Preprocessing and Model Training

To prepare the ChestX-ray14 dataset for training, we applied a series of preprocessing steps to enhance model performance and generalization:

- Image transformations:
  - **Grayscale conversion:** Converted images to 3-channel grayscale to match the input format expected by most pre-trained models.
  - **Random horizontal flip:** Introduced variability in image orientation, helping the model generalize better to unseen cases.
  - **Random rotation (up to 15°):** Added rotational invariance to improve robustness to slight changes in patient positioning.
  - **Resizing to** 224×224**:** Standardized input dimensions for compatibility with deep learning models and to reduce computational load.
  - **Normalization:** Standardized pixel intensity values using ImageNet mean and standard deviation to stabilize and accelerate training.
- **Multi-label encoding:** Each image may be associated with multiple disease labels. We used multi-label binarization to convert these into a format suitable for training.
- **Patient-level split:** To prevent data leakage, all images from a given patient were assigned to only one of the data splits. We used a 70% training, 10% validation, and 20% test split.

These preprocessing steps ensured that input data was consistent, diverse, and leakage-free for effective model training and evaluation.

Enhancing Preprocessing with Gamma Correction

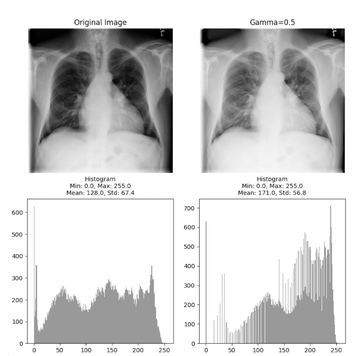


Figure 4. Gamma-corrected chest X-ray sample.

To enhance contrast sensitivity and model robustness, we applied **gamma correction** as an additional data augmentation technique. This non-linear transformation helps simulate varying X-ray exposures and improves the model's ability to detect fine-grained features — particularly in under-represented conditions [2]. Each model was trained both with and without gamma correction, and the best-performing variant was selected for the final ensemble. Gamma correction led to measurable improvements in AUROC, especially for minority classes:

- **ConvNeXt:** Highest overall performance – 83.40% AUROC
- **MaxViT – Edema (2.05%): 90.35% AUROC**
- **Swin – Pleural Thickening (3.02%): 82.08% AUROC**

Singular Model Results

Model	AUROC	4-Fold CV					
		Mean	SD	Fold 1	Fold 2	Fold 3	Fold 4
VGG-19	0.8040	0.8007	0.0055	0.8022	0.7928	0.7997	0.8007
DenseNet121	**0.8302**	0.8289	0.0016	0.8264	0.8299	0.8287	0.8306
MaxViT	**0.8331**	0.8360	0.0020	0.8345	0.8360	0.8340	0.8392
CoAtNet	**0.8326**	0.7721	0.0921	0.8259	**0.6126**	0.8250	0.8251
ConvNeXt	**0.8323**	0.8344	0.0007	0.8346	0.8332	0.8351	0.8347
Swin	0.8299	0.8213	0.0024	0.8289	0.8291	0.8239	0.8299

Table 1. Performance of individual models with 4-Fold CV: mean AUROC, standard deviation, and per-fold results.

Ensembling Methods and Analysis

To boost performance, we ensemble the outputs of individual models using two strategies:

1. Weighted Average Ensemble: We combine model outputs using a weighted sum:

$$\hat{y}_i = \sum_{k=1}^{K} w_k \cdot p_{i,k}$$

where $\hat{y}_i$ is the final ensemble prediction for sample i, w_k is the weight for model k, and $p_{i,k}$ is the prediction of model k for sample i.

2. Differential Evolution (DE) + Greedy Selection Weighted Ensemble: DE optimizes the weights $w_1, w_2, \ldots, w_K$ to maximize the mean AUROC across all samples:

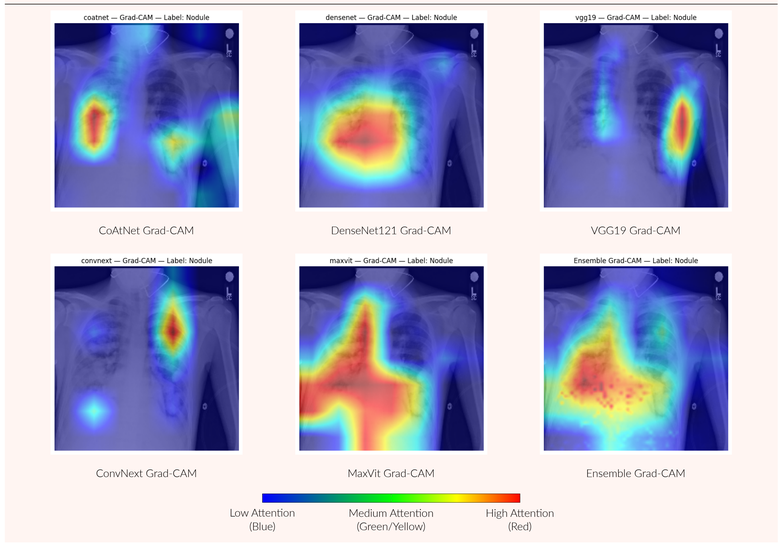
$$\max_{w_1,w_2,\ldots,w_K} \frac{1}{N} \sum_{i=1}^{N} \text{ROC-AUC}(y_{\text{true},i}, \hat{y}_i)$$

DE optimizes weights iteratively as greedy selection incrementally builds the ensemble.

Ensemble Method	AUROC	Competitive Best
Weighted Average	0.8562	0.8532
Differential Evolution	0.8565	0.8543

Table 2. AUROC scores of ensemble methods on the ChestX-ray14 dataset compared to competitive baselines [1].

Heatmap Interpretability on Ensemble Grad-CAM



References

[1] S. M. N. Ashraf, Md. A. Mamun, H. Md. Abdullah, and Md. G. Alam. "Synthensemble: A fusion of CNN, vision transformer, and hybrid models for multi-label chest X-ray classification." *2023 International Conference on Computer and Information Technology (ICCIT)*, pp. 1–6. https://doi.org/10.1109/iccit60459.2023.10441433

[2] S. I. Maiyanti et al. "Rotation-gamma correction augmentation on CNN-dense block for soil image classification." *Applied Computer Science*, vol. 19, no. 3, pp. 96–115, 2023. https://doi.org/10.35784/acs-2023-27

[3] K. L. Ong et al. "MaxMViT-MLP: Multiaxis and multiscale vision transformers fusion network for speech emotion recognition." *IEEE Access*, vol. 12, pp. 18237–18250, 2024. https://doi.org/10.1109/access.2024.3360483

[4] V. Tynchenko, S. Kurashkin, and V. Kukartsev. "Application of convolutional neural networks to determine induction soldering process technological stages." *European Proceedings of Computers and Technology*, vol. 3, pp. 210–221, 2023. https://doi.org/10.15405/epct.23021.26

[5] X. Wang et al. *ChestX-ray14 (v1)* [Data set]. National Institutes of Health Clinical Center, 2017.