Chest X-Ray Image Classification
- using Deep Learning approach

Project Report

Contents

Abstract

Chest X-rays are a vital diagnostic tool, but interpreting them for multiple co-occurring diseases is challenging, especially when some conditions are rare. In this project, we developed a multi-label classification model to improve detection accuracy across both common and rare thoracic pathologies in the NIH ChestX-ray14 dataset. Our approach compared a baseline convolutional neural network (CNN), DenseNet121, and EfficientNetB0 against a new Dense–Efficient Attention-Fusion Network, which combines DenseNet121 and EfficientNetB3 features through a cross-attention mechanism. All models were trained using Focal Loss to address severe class imbalance. The fusion model achieved a macro-averaged Area Under the Receiver Operating Characteristic Curve (ROC-AUC) of 0.8156 on the test set, with marked improvements for underrepresented diseases, raising Hernia AUC from 0.53 (EfficientNetB0) to 0.94 and Emphysema from 0.77 to 0.93. Results show that integrating complementary feature extractors with attention and targeted loss functions enhances performance on both frequent and rare conditions. Incorporating uncertainty estimation using Monte Carlo (MC) Dropout technique to the proposed fusion model further enhanced the credibility of the model. This architecture offers a step toward more reliable, balanced, and interpretable computer-aided chest X-ray diagnosis, supporting radiologists in clinical decision-making.

Introduction

Chest radiography is the most widely ordered imaging test in clinical practice, yet subtle findings are still missed in up to one-quarter of first reads. Errors are most common when several diseases overlap on the same film or when a rare condition appears only a handful of times a year. Computer-aided diagnosis systems based on deep learning have begun to ease this burden, but two practical gaps remain. First, most models focus on a single disease or a small set of common labels, leaving rare findings under-served. Second, they seldom tell the reader how confident they are, making it hard to know when a second opinion is needed.

This work tackles both issues by building a multi-label chest-X-ray classifier that (i) treats fourteen thoracic conditions at once, (ii) improves recognition of the least frequent classes, and (iii) provides calibrated confidence scores. We start with three established baselines: a simple CNN, DenseNet121, and EfficientNetB0, then introduce a *Dense–Efficient Attention-Fusion Network*. To cope with severe label imbalance, we train every model with Focal Loss. To expose predictive uncertainty, we insert Monte-Carlo dropout in the classification head and average 25 stochastic passes per image.

All experiments use the NIH ChestX-ray14 dataset, which contains 1,12,120 frontal X-rays from 30,805 patients, with 14 distinct pathology labels.

Our main contributions are:

- A cross-attention fusion architecture that combines DenseNet and EfficientNet features for richer representations.

- A training scheme that joins Focal Loss and Monte-Carlo dropout, boosting rare-class recall while exposing model confidence.
- A systematic comparison with three baselines on the full multi-label ChestX-ray14 task, using identical splits and hyper-parameters.

Background

As computational power and confidence in artificial intelligence (AI) applications continue to grow, AI-driven medical image analysis has seen rapid adoption. Deep learning architecture, an advanced subfield of artificial intelligence, has delivered promising results in medical imaging. Numerous studies now demonstrate that deep learning models can accurately diagnose diseases from chest X-ray images. The foundational works like deep learning architecture named CheXNet, which is a 121-layer dense convolutional neural network-based model, achieved radiologist-level pneumonia detection (1). Wang et al. (5) performed disease detection from chest X-ray images using the ChestX-ray8 dataset prepared at hospital scales. There are data on 8 different diseases in the data set; these are atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia and pneumothorax. The authors achieved 63.9% success in AlexNet architecture, 63.9% in GoogLeNet architecture, 62.5% in VGGNet-16 architecture, and finally 69.6% in ResNet-50 architecture.

Recent multi-label chest X-ray classification research leverages advanced CNN architectures such as DenseNet, ResNet, and EfficientNet-B3, achieving high accuracy on public datasets like CheXpert and ChestX-ray14 (2). To address the critical need for reliability, many approaches integrate Monte Carlo Dropout (MC-Dropout) for uncertainty estimation—e.g., UA-ConvNet fine-tuned on EfficientNet-B3 reports G-mean ~ 98% with meaningful confidence intervals (3). Large-scale studies (e.g., Whata et al., 2023) show that ensemble MC-Dropout and Bayesian networks significantly improve calibration and uncertainty-aware metrics in multi-class settings (4).

Building on prior research, we explored some CNN architectures and experimented with a fusion network that integrates DenseNet, EfficientNet, and a cross-attention mechanism.

Approach

This section provides a detailed description of the dataset characteristics employed during the training, testing, and validation phases and the pre-processing steps performed. It presents comprehensive information on the deep learning models utilized as well as other techniques used in our experiments.

Data Pre-processing

Our primary dataset for this project was the **NIH ChestX-ray14 dataset** (6), which contains 112,120 chest X-ray images of 30,805 different patients, with 14 distinct pathology labels. Every patient has at least one chest X-ray image, and some patients have multiple images. *Figure 1* shows the distribution of diseases in the dataset.

The diseases found in the data are as follows: **atelectasis, cardiomegaly, consolidation, edema, effusion, emphysema, fibrosis, hernia, infiltration, mass, nodule, pleural thickening, pneumonia, pneumothorax**. The images in the dataset are 8-bit black and white images. To manage computational resources, a subset of the data, representing 50% of the total patients, was randomly sampled for training and evaluation.

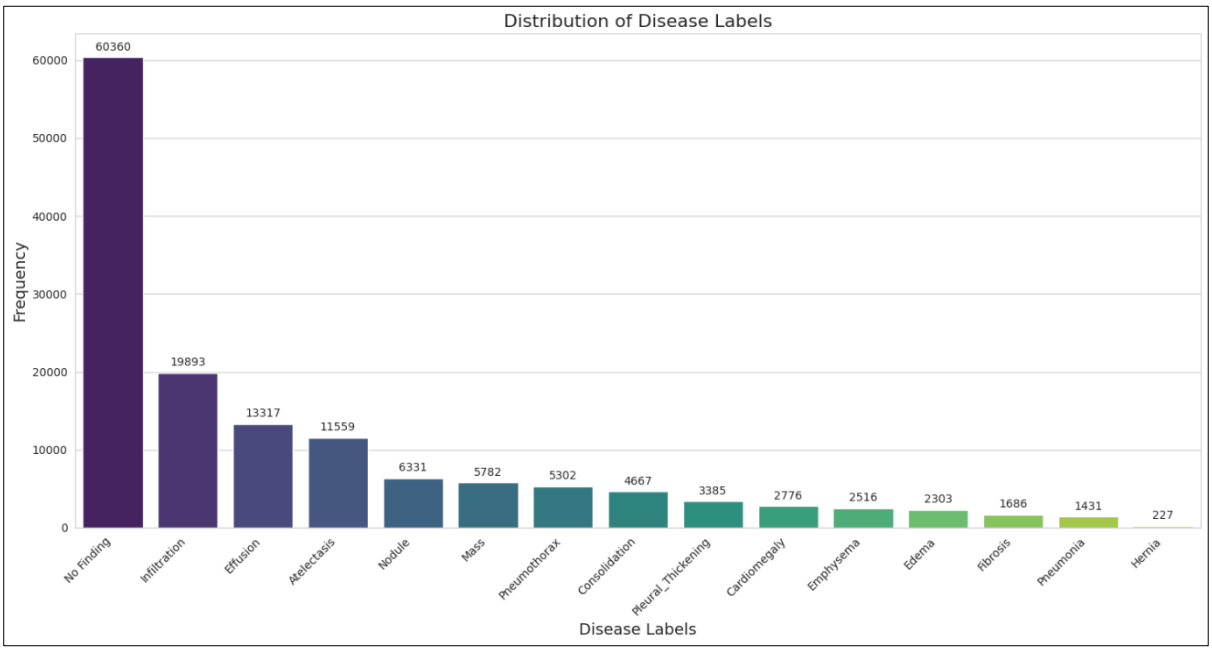

Figure1: Distribution of 14 pathologies in the dataset. The label "No Finding" suggests images where label is missing.

The preprocessing pipeline involved several key steps to prepare the data for our multi-label classification task. The dataset comes with a publicly available CSV file, which served as a data matching tool that contains image names and disease names. A new column "full_path" was created and populated with the relative path of the images, to be used for loading the images. Next, the raw text labels were converted into a multi-hot encoded vector. Each of the 14 pathologies (e.g., Cardiomegaly, Hernia, Effusion) was represented by a binary column. The "No Finding" label was treated as the absence of all 14 pathologies. To ensure that the distribution of rare label combinations was preserved across datasets, we used **Iterative Stratification**. The data was split into a 70% training set, a 10% validation set, and a 20% test set. This stratified approach is crucial for preventing model bias and ensuring reliable evaluation in a multi-label context. Finally, to improve model generalization and reduce overfitting, a series of random augmentations were applied to the training images. This included random horizontal flips, rotations (10%), zooming (10%), and brightness adjustments (10%). All images were resized to a standard 224x224 pixel resolution.

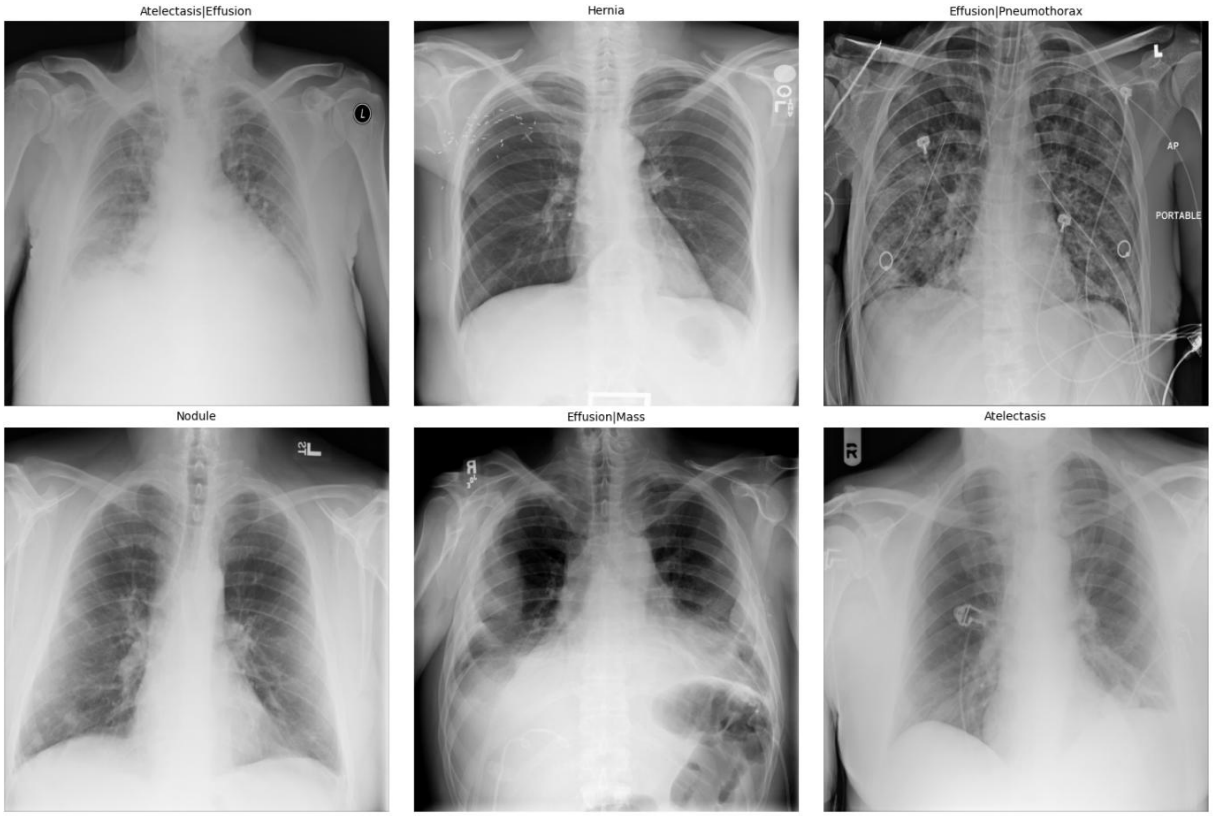

Figure 2: Sample Images from NIH ChestX-ray14 dataset

Model Architectures and Training

We evaluated few architectures to establish a performance baseline and demonstrate the effectiveness of our proposed model. We compared four architectures:

- A **simple baseline CNN**
- **DenseNet121** pretrained on ImageNet as the backbone, followed by global average pooling, an MC Dropout layer, a 512-unit dense layer with dropout, and a 14-unit output layer.
- **EfficientNetB0** pretrained on ImageNet as the backbone, followed by global average pooling, an MC Dropout layer, a 512-unit dense layer with dropout, and a 14-unit output layer.
- Our proposed fusion model **combining DenseNet121 and EfficientNetB3 with a cross-attention mechanism**.

Figures 3a-c shows the model summary of the deep neural network architectures. For models using DenseNet121 and EfficientNetB0 as backbone only the summary of trainable parameters on the custom layers are being shown.

Model: "Simple CNN"

Layer (type)	Output Shape	Param #
input_layer_1 (InputLayer)	(None, 224, 224, 3)	0
conv2d (Conv2D)	(None, 222, 222, 32)	896
max_pooling2d (MaxPooling2D)	(None, 111, 111, 32)	0
conv2d_1 (Conv2D)	(None, 109, 109, 64)	18,496
max_pooling2d_1 (MaxPooling2D)	(None, 54, 54, 64)	0
conv2d_2 (Conv2D)	(None, 52, 52, 128)	73,856
max_pooling2d_2 (MaxPooling2D)	(None, 26, 26, 128)	0
global_average_pooling2d (GlobalAveragePooling2D)	(None, 128)	0
dense (Dense)	(None, 256)	33,024
dense_1 (Dense)	(None, 14)	3,598

Total params: 129,870 (507.30 KB)
Trainable params: 129,870 (507.30 KB)
Non-trainable params: 0 (0.00 B)

Figure 3a. Model Summary of Simple CNN architecture

----------Custom Layer Summary: ----------
Model: "DenseNet121"

Layer (type)	Output Shape	Param #
input_layer_2 (InputLayer)	(None, 224, 224, 3)	0
densenet121 (Functional)	(None, 7, 7, 1024)	7,037,504
global_average_pooling2d (GlobalAveragePooling2D)	(None, 1024)	0
dropout (Dropout)	(None, 1024)	0
dense (Dense)	(None, 512)	524,800
dropout_1 (Dropout)	(None, 512)	0
predictions (Dense)	(None, 14)	7,182

Total params: 7,569,486 (28.88 MB)
Trainable params: 531,982 (2.03 MB)
Non-trainable params: 7,037,504 (26.85 MB)

----------Custom Layer Summary: ----------
Model: "EfficientNetB0"

Layer (type)	Output Shape	Param #
input_layer_2 (InputLayer)	(None, 224, 224, 3)	0
efficientnetb0 (Functional)	(None, 7, 7, 1280)	4,049,571
global_average_pooling2d (GlobalAveragePooling2D)	(None, 1280)	0
dropout (Dropout)	(None, 1280)	0
dense (Dense)	(None, 512)	655,872
dropout_1 (Dropout)	(None, 512)	0
predictions (Dense)	(None, 14)	7,182

Total params: 4,712,625 (17.98 MB)
Trainable params: 663,054 (2.53 MB)
Non-trainable params: 4,049,571 (15.45 MB)

Figures: 3b. Model Summary of custom layer of DenseNet121; 3c. Model Summary of custom layer of EfficientNetB0

Our proposed model is the one with fusion architecture: **Dense-Efficient Attention-Fusion Network**. This model was designed to create a highly expressive feature representation by combining the strengths of two distinct, powerful backbones: **DenseNet121** and **EfficientNetB3.** DenseNet121 and EfficientNetB3 were used as parallel feature extractors. We hypothesized that their different architectural designs would capture complementary features. Instead of late-stage ensemble averaging, we extracted feature maps from intermediate layers of both backbones. These maps were then aligned to a common channel dimension (256) **using 1x1 convolutions** and resized to match spatially before being concatenated. The fused feature map was passed through a custom **Cross-Attention layer**. This self-attention mechanism refines the combined features by learning to weigh the importance of different spatial locations and inter-channel relationships, effectively amplifying the most salient signals for diagnosis. A **global average pooling layer**, a **dropout layer** (rate of 0.5), and a **final dense layer with a sigmoid activation** function produced the 14-label probability output. Total parameters for this network are 10,638,528 (40.58 MB); **Trainable parameters being 10,521,273 (40.14 MB).** *Figure 4* shows an overall architecture of the fusion network.

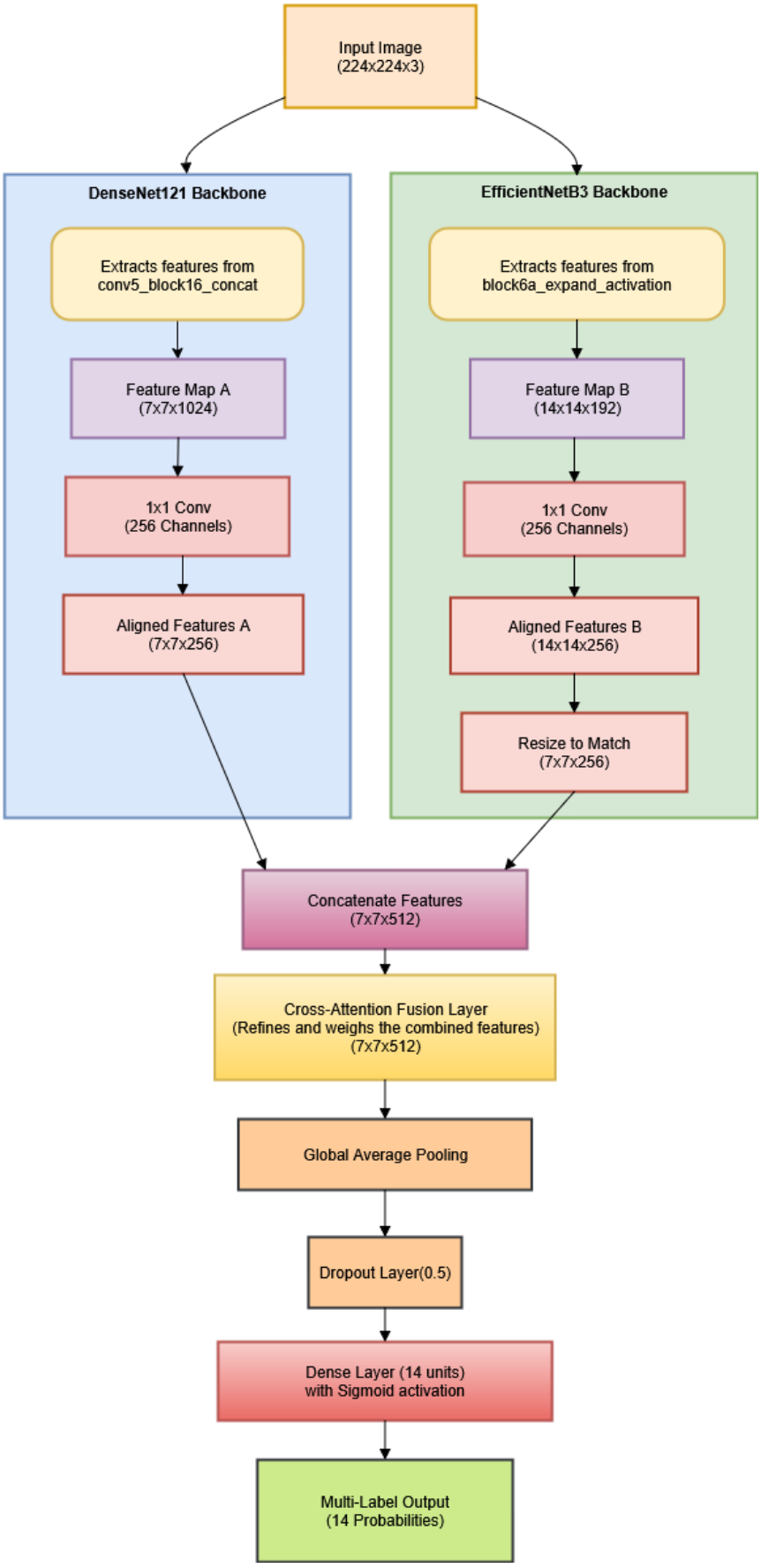

Figure 4: Overview of Dense-Efficient Attention-Fusion Network architecture

Addressing Class Imbalance using Focal Loss Function

Another key objective of our project was to address the severe class imbalance inherent in the dataset. We selected **Focal Loss** as our primary method to tackle this challenge. Focal Loss dynamically down-weights the loss contribution of easy, well-classified examples (often the negative samples of a rare disease). This forces the model to focus its training efforts on hard-to-classify, infrequent positive samples, thereby improving its ability to recognize rare pathologies.

Mathematically, for a single example with ground-truth label y ∈ {0,1} and model-estimated probability p=σ(x) for the positive class, we define

$$p_t = \begin{cases} p, & y = 1 \\ 1 - p, & y = 0 \end{cases}$$

Then the Focal Loss is

$$\mathcal{L}_{FL}(p_t) = -\alpha(1 - p_t)^{\gamma} \log(p_t)$$

where:

α ∈ [0,1] balances the importance of positive versus negative examples,

γ ≥ 0 is the focusing parameter: larger γ increases the down-weighting of well-classified examples.

By modulating the loss with $(1 - p_t)^{\gamma}$, Focal Loss reduces the relative loss for easy examples (where p_t is high) and effectively concentrates the gradient updates on harder, misclassified samples. This property makes it particularly effective for tasks such as chest X-ray disease detection, where certain pathologies are rare and can otherwise be overlooked by conventional loss functions.

Incorporation of Uncertainty Estimation using Monte Carlo Dropout

Establishing trust in clinical machine learning requires more than single-point predictions - an associated measure of confidence is what is preferred. To address this, we used MC Dropout to estimate model uncertainty.

MC Dropout is a Bayesian approximation technique that repurposes standard dropout layers—typically used for regularization during training—into a tool for uncertainty quantification at inference time. Instead of deactivating dropout during testing, we perform multiple forward passes on the same input image with dropout enabled. This process generates a distribution of predictions for each class. The mean of this distribution serves as the final predictive probability, while the variance (or standard deviation) serves as a measure of the model's uncertainty. A high variance signifies that the model is uncertain about its prediction, alerting a clinician that the result may be unreliable and requires closer human inspection. This approach provides not just a classification but also a confidence level, making the diagnostic tool more transparent and trustworthy.

Results

Experimental Setup

We utilized the NIH ChestX-ray14 dataset for all experiments. To ensure efficient training and iteration, we worked with a randomly sampled subset representing 50% of the unique patients. The data was partitioned using Iterative Stratification into training (70%), validation (10%), and test (20%) sets to maintain the complex multi-label distributions.

All models were trained using the same training and evaluation framework, on 224x224 pixel images with a batch size of 32. To address the significant class imbalance in the dataset, **Focal Loss** was used as the loss function across all experiments, with hyperparameters α=0.25 and γ=2.0. Table 1 shows a list of hyperparameters used.

Except simple CNN model, others were trained using a two-phase approach. Initially, only the custom classification head was trained, while all convolutional layers remained frozen. This head comprised global average pooling, dropout with Monte Carlo sampling (rate = 0.4), a 512-unit ReLU layer with dropout, and a final 14-unit sigmoid output. We employed Adam with an initial learning rate of 1×10^{-3}, optimizing a focal loss to mitigate class imbalance and monitoring multi-label AUC. After early stopping and checkpointing on validation AUC, we unfroze the entire DenseNet/EfficientNet and fine-tuned with a reduced learning rate (1×10^{-5}). The same training process was used for our proposed model; initial training of the fusion layers followed by fine-tuning of the entire network.

Hyperparameter	Value
Optimizer	Adam
Metrics	AUC
Epochs	25
Initial learning rate	1e−3
Minimum learning rate	1e−5
Alpha (Focal Loss)	0.25
Gamma (Focal Loss)	2.0
Dropout Rate	0.5
Batch size	32

Table 1: Hyper-parameters used in the training phase

The primary performance metric was the macro-averaged **Area Under the Receiver Operating Characteristic Curve (ROC-AUC)**, evaluated on the held-out test set. This metric is ideal for imbalanced multi-label tasks as it measures the model's ability to discriminate between classes.

Model Performance and Evaluation

The experiments showed that the Simple CNN as well as the standalone pre-trained models, DenseNet121 and EfficientNetB0, served as strong baselines. While they achieved reasonable performance on common pathologies, they struggled significantly with rare diseases, even with the aid of Focal Loss. *Figure 5a*, *5b, 5c* show the ROC curves for each pathology on the test set for Simple CNN, DenseNet121 and EfficientNetB0, respectively. Notably, the AUC

for the rare disease **Hernia** was only **0.68** for DenseNet121 and **0.53** for EfficientNetB0, the latter being only slightly better than random chance.

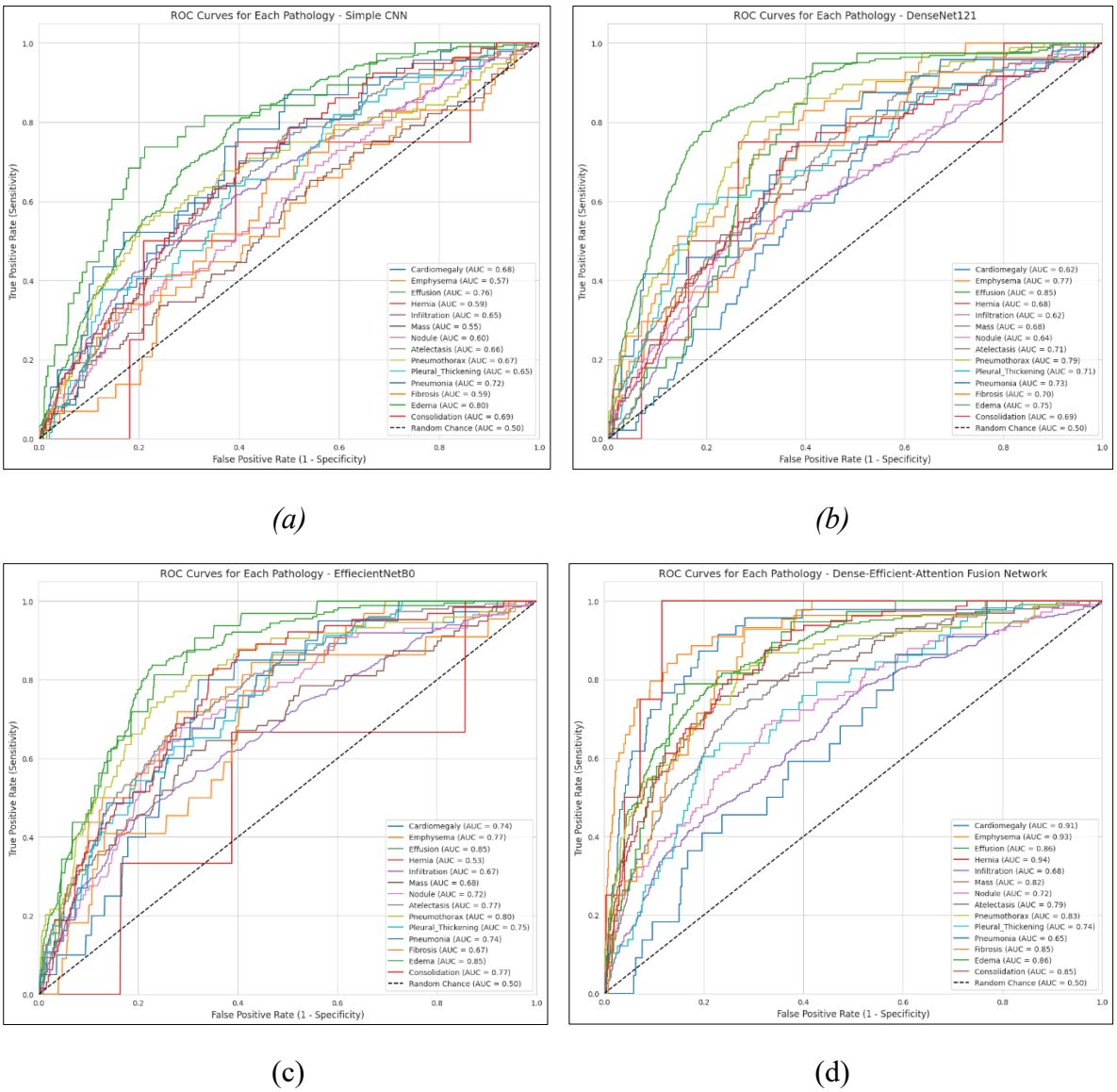

Figure 5: ROC Curves on Test Set for (a) Simple CNN (b) DenseNet121 (c) EfficientNetB0 (d) Dense-Efficient Attention-Fusion Network

The fusion architecture demonstrated a substantial improvement over the baseline models. The final macro-averaged test AUC was **0.8156**. The model's ability to diagnose rare diseases saw the most improvement, validating our fusion and attention-based approach. *Figure 5d* shows the ROC curves for Dense-Efficient Attention-Fusion Network. The performance on **Hernia** increased to an AUC of **0.94**, a massive gain over the standalone models. As shown in *Table 2*, significant gains were also observed for other challenging classes like Emphysema and Fibrosis. However, the model found it more challenging to classify pathologies known for their subtle or overlapping visual features, such as: Pneumonia (0.65), Infiltration (0.68), Nodule (0.72).

Pathology	Simple CNN	DenseNet121	EfficientNetB0	Fusion Network
Cardiomegaly	0.68	0.62	0.74	**0.91**
Emphysema	0.57	0.77	0.77	**0.93**
Effusion	0.76	0.85	0.85	0.86
Hernia	0.59	0.68	0.53	**0.94**
Infiltration	0.65	0.62	0.67	0.68
Mass	0.55	0.68	0.68	**0.82**
Nodule	0.59	0.64	0.72	0.72
Atelectasis	0.67	0.71	0.76	**0.79**
Pneumothorax	0.67	0.79	0.80	**0.83**
Pleural Thickening	0.65	0.71	0.75	0.74
Pneumonia	0.72	0.73	0.74	0.65
Fibrosis	0.59	0.69	0.67	**0.85**
Edema	0.79	0.75	0.85	**0.86**
Consolidation	0.69	0.69	0.77	**0.85**
Macro-Averaged AUC	0.6558	0.7096	0.7363	**0.8156**

Table 2: Comparison of Test AUC Scores for each pathology

Figure 6 displays four graphs comparing the **training loss** and **validation loss** over 25 epochs for the four neural network models. The goal during training is to minimize the loss function, and these plots are crucial for diagnosing model performance, specifically its ability to generalize to new, unseen data.

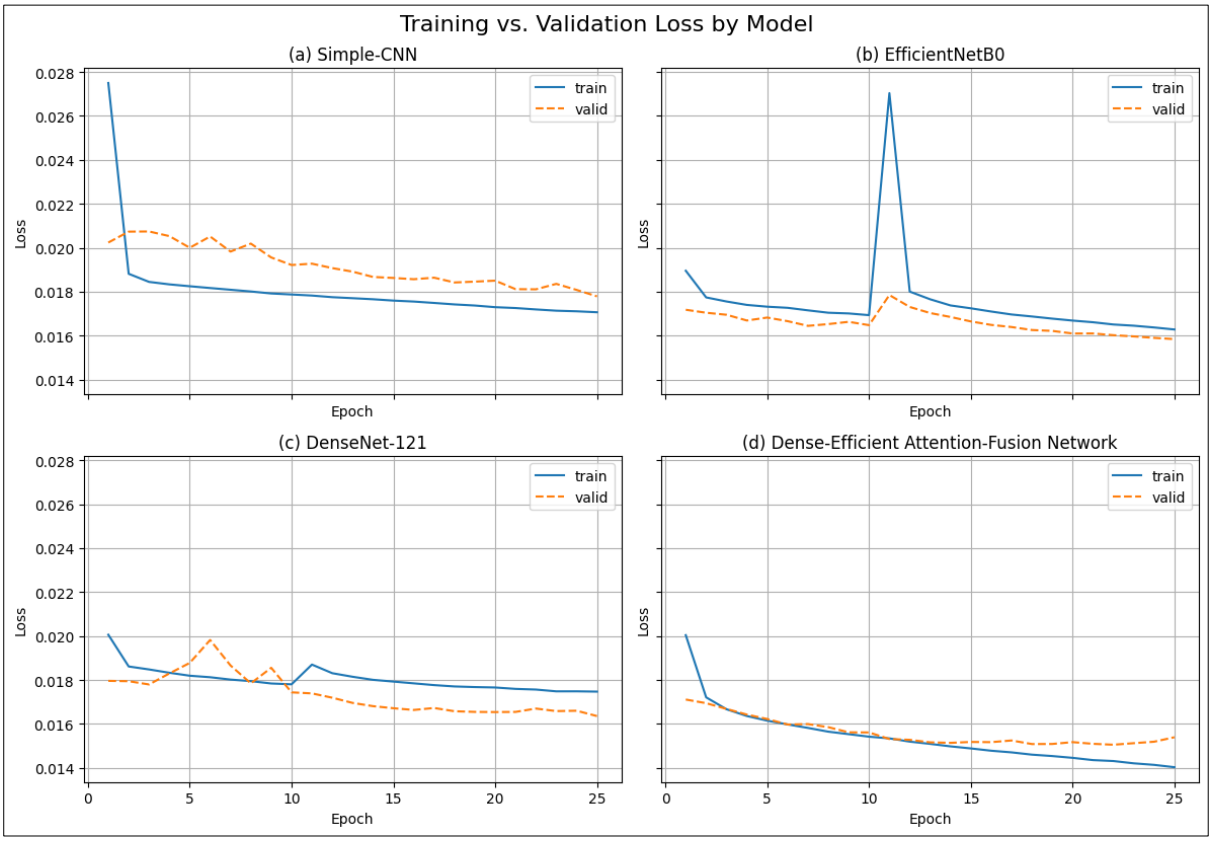

Figure 6: Training vs Validation Loss over 25 epochs for each model

A visual analysis of the loss curves shows that the Dense-Efficient Attention-Fusion Network (d) is the superior model. It demonstrates stable training, the lowest final validation loss, and excellent generalization capabilities. In contrast, the other models exhibit various issues: the Simple-CNN (a) shows moderate overfitting, the EfficientNetB0 (b) suffers from a notable instability spike, and the DenseNet-121 (c) fails to generalize effectively and shows significant instability.

The objective of incorporating uncertainty estimation was successfully achieved using the MC Dropout technique. By performing **100 stochastic forward passes** on a sample test image with dropout enabled during inference, the model generated a distribution of outcomes, allowing for the calculation of prediction confidence. For an example test image with a true label of **'Atelectasis'**, the analysis produced the following results:

- Prediction: The model's mean predictive probability for the 'Atelectasis' class was **30.29%,** which was the highest probability among all classes, leading to a correct classification.
- Uncertainty Score: The average uncertainty (standard deviation across all class probabilities) was **0.0169**. This very low value falls well below a typical uncertainty threshold (0.1 in our case), indicating high model confidence in its prediction for this specific image.

Figure 6 shows the corresponding histogram of the 100 prediction probabilities for the 'Atelectasis' class, showing a tight, narrow distribution centered around the mean. This visually confirms the low variance and high certainty of the model's output.

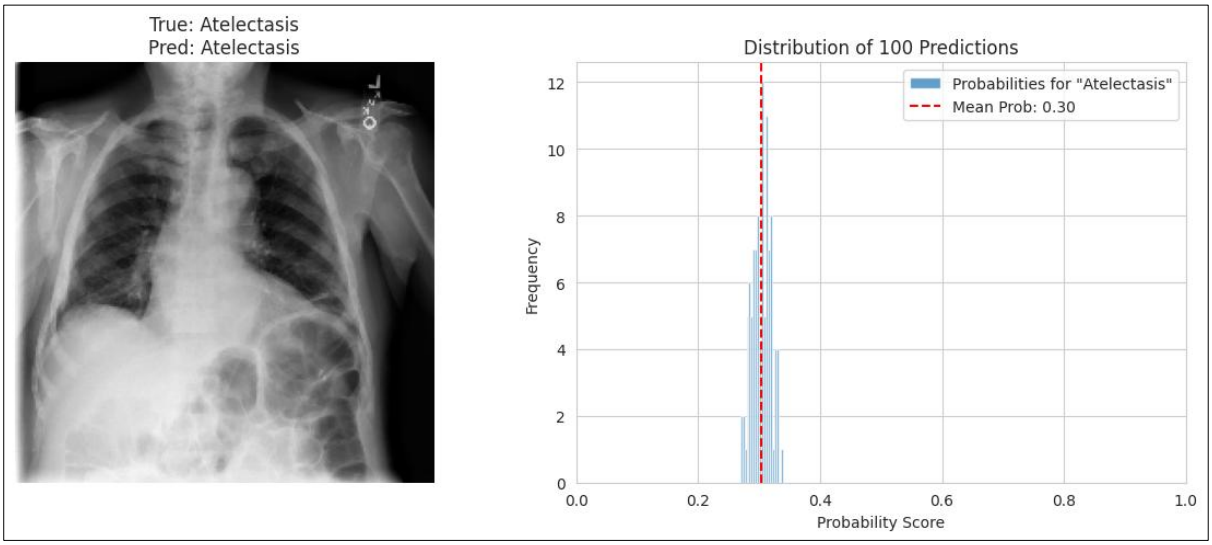

Figure 7: Model output of uncertainty estimation on a test image

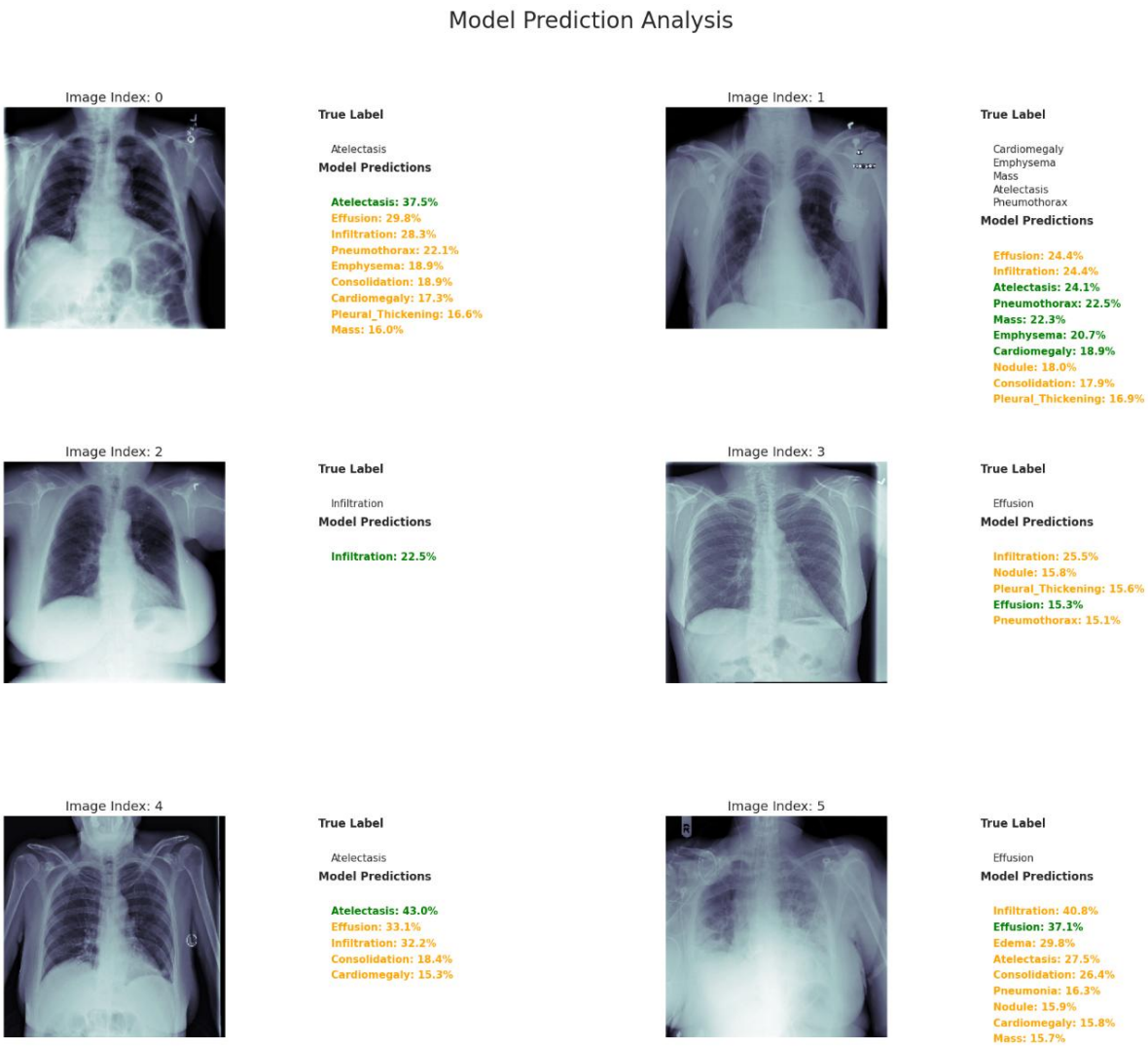

Figure 8: Training vs Validation Loss over 25 epochs for each model

Discussion

The results demonstrate that the proposed Dense–Efficient Attention-Fusion Network substantially outperforms both the simple CNN and the two standalone ImageNet backbones (DenseNet121 and EfficientNetB0) across most pathologies, particularly the rare classes. By integrating feature maps from two complementary architectures via cross-attention, the fusion model leverages the strengths of both DenseNet's fine-grained feature propagation and EfficientNet's efficient scaling. The most remarkable improvements were observed in the detection of rare pathologies. For instance, ROC-AUC for Hernia surged from as low as 0.53 in the EfficientNetB0 model to 0.94 in our fusion model. Similarly, the AUC for Emphysema and Fibrosis saw dramatic increases, reaching 0.93 and 0.85, respectively. These improvements validate our hypothesis that combining diverse feature extractors can better capture subtle radiographic patterns that single backbones may miss. The effectiveness of the model is also reflected in the stable convergence behaviour; the fusion network exhibited the lowest validation loss and most consistent training curves among all architectures *(Figure 6d)*.

Training with Focal Loss proved critical for addressing severe label imbalance. By down-weighting well-classified majority examples, the model focused learning on hard, minority-class samples, boosting recall on hard-to-classify pathologies. While Focal Loss improved the performance of baseline models, its combination with the powerful feature representation of the fusion network yielded the most significant gains.

Monte Carlo Dropout successfully quantified predictive uncertainty, adding an essential layer of confidence estimation. On example Atelectasis cases, the low standard deviation (0.0169) across 100 stochastic forward passes confirmed that the model's high-probability predictions are internally consistent. Such uncertainty estimates can guide clinicians toward cases warranting further review, enhancing trust in automated systems.

However, challenges remain for pathologies with inherently subtle or overlapping visual markers like Pneumonia (AUC 0.65) and Infiltration (AUC 0.68), suggesting that feature extraction for these conditions requires further refinement. Figure 7 shows the output of model predictions on few test images, which also reflects scope for improvement in model architecture. Moreover, our experiments used a 50% patient subset for efficiency; future work should validate scalability on the full ChestX-ray14 corpus.

Conclusion

This project successfully developed and validated a novel deep learning architecture, the Dense-Efficient Attention-Fusion Network, for the multi-label classification of 14 thoracic pathologies from chest X-rays. Our model demonstrates a significant improvement in diagnostic accuracy over established baselines, achieving a macro-averaged ROC-AUC of 0.8156 on the test set. By integrating a training regimen that combines Focal Loss to counteract severe class imbalance and Monte Carlo Dropout to estimate predictive uncertainty, we have addressed two critical gaps in computer-aided chest X-ray diagnosis. This approach not only boosts the recall of underrepresented conditions but also provides a measure of model confidence, enhancing the system's reliability and interpretability for clinical use. Our results, particularly the dramatic performance increase for pathologies like Hernia, Emphysema, and Fibrosis, highlight the value of our methodology.

Future work should focus on refining the model's ability to distinguish pathologies with subtle visual features, such as Pneumonia and Infiltration. Further exploration of different fusion techniques and attention mechanisms could yield additional performance gains.

References

1. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, Ding D, Bagul A, Langlotz C, Shpanskaya K, Lungren MP, Ng AY. 2017. CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning. (https://arxiv.org/abs/1711.05225)

2. Liz, Helena & Huertas-Tato, Javier & Sánchez-Montañés, Manuel & Del Ser, Javier & Camacho, David. (2022). Deep learning for understanding multilabel imbalanced Chest X-ray datasets. 10.48550/arXiv.2207.14408.
3. Mahesh Gour, Sweta Jain, Uncertainty-aware convolutional neural network for COVID-19 X-ray images classification, Computers in Biology and Medicine, Volume 140, 2022, 105047, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2021.105047 .

4. Whata A, Dibeco K, Madzima K and Obagbuwa I (2024) Uncertainty quantification in multi-class image classification using chest X-ray images of COVID-19 and pneumonia. Front. Artif. Intell. 7:1410841. doi: 10.3389/frai.2024.1410841 (https://doi.org/10.3389/frai.2024.1410841)

5. 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," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 3462-3471, doi: 10.1109/CVPR.2017.369.

6. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. IEEE CVPR 2017