

Benchmarks and Improvements on Multi-Label Chest X-ray Classification Using a Lightweight CNN Architecture

Ahmad Hamadi¹, Jared Paul¹, Gurnoor Bal¹, Hamza Issa¹, Harrison Chiu¹

¹Department of Computing and Software, McMaster University,
Hamilton, Ontario, Canada

Email: hamada11@mcmaster.ca, paulj20@mcmaster.ca,
balg3@mcmaster.ca, issah3@mcmaster.ca, chiuh6@mcmaster.ca

Capstone Research Paper

Abstract

Abstract Chest radiography is a foundational imaging tool in clinical diagnostics. In this work, we present a lightweight convolutional neural network for multi-label classification of chest X-rays. We explore different backbone architectures, loss functions, and data balancing strategies to optimize classification performance on the NIH ChestX-ray14 dataset. Our findings demonstrate that with careful tuning, compact models can achieve performance comparable to larger networks while remaining efficient, interpretable, and suitable for deployment in resource-constrained healthcare environments.

1 Introduction

Chest radiography (CXR) remains one of the most essential imaging techniques in modern medicine, widely used for diagnosing thoracic abnormalities such as pulmonary or cardiac conditions. However, interpreting CXRs is inherently difficult due to overlapping anatomical structures, subtle pathological findings, and variability in radiologist interpretation. These challenges are further amplified by rising patient volumes and limited radiology staff, driving the need for intelligent diagnostic support tools [1].

Recent advancements in deep learning have enabled artificial intelligence (AI) models to assist or exceed human-level performance in radiographic analysis. Large-scale public datasets like NIH ChestX-ray14 [1] have made it possible to train convolutional neural networks (CNNs) for thoracic disease classification. Despite this progress, real-world deployment still faces several key obstacles: highly imbalanced label distributions [2], label noise from NLP-extracted radiology reports [3], and the computational demands of large models [4].

Our capstone project addresses these concerns by developing a lightweight, inter-

pretable model for multi-label CXR classification using MobileNetV2. By carefully optimizing our training process through class rebalancing, data augmentation, dynamic thresholding, and a custom margin-based loss function, we demonstrate that smaller models can achieve competitive performance [5].

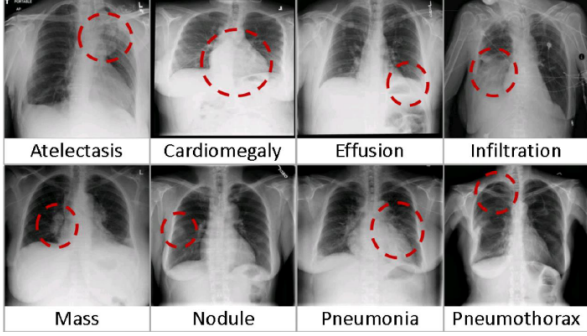


Figure 1.1. Example chest X-rays from the ChestX-ray14 dataset labeled with common thoracic diseases. Red circles highlight affected regions.

2 Background and Prior Work

2.1 Medical Imaging and Chest X-rays

Chest radiography is one of the most accessible and widely used imaging tools in global healthcare. It plays a key role in detecting and monitoring thoracic conditions such as pneumonia, effusion, and cardiomegaly [1]. However, CXR interpretation remains a difficult task, even for trained radiologists, due to overlapping anatomical structures, low-contrast findings, and significant inter-reader variability [1]. These factors highlight the growing need for AI-based tools to support interpretation, especially in settings with limited expert availability.

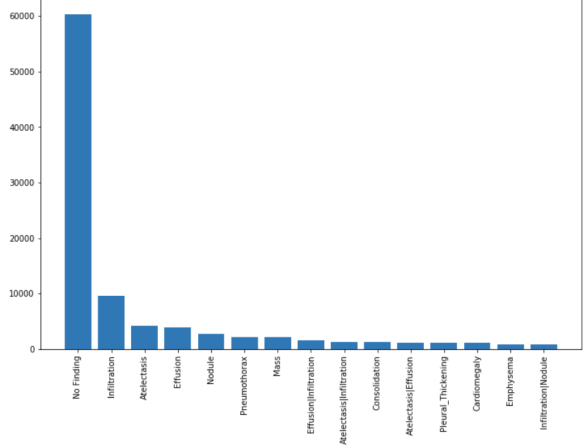


Figure 2.1. Sample CXR images from ChestX-ray14 showing multiple thoracic conditions with bounding-box annotations [1].

2.2 Deep Learning in Radiology

Deep learning, particularly with Convolutional Neural Networks (CNNs), has revolutionized image classification across domains, including medical imaging. CNNs automatically extract spatial features from pixel data, learning hierarchical representations without hand-crafted features. In radiology, architectures like AlexNet, VGG, and ResNet have demonstrated strong results on chest X-ray benchmarks [1, 2].

However, large architectures come with significant computational demands. To address this, researchers have developed lightweight models like MobileNetV2, which replaces standard convolutions with depth-wise separable convolutions, drastically reducing parameter count and improving inference speed without major performance tradeoffs [5].

2.3 Multi-Label Classification Challenges

Chest X-ray classification differs from traditional single-label classification problems. It is multi-label, meaning each image can contain multiple co-occurring conditions. This introduces several unique challenges:

Class imbalance is a significant problem. For example, “Infiltration” and “Effusion” appear in tens of thousands of images, while rarer conditions like “Hernia” are present in fewer than 200 samples in the NIH dataset [1]. Training without adjustments often causes models to over-predict dominant labels and ignore rare ones [2].

Label noise is another critical issue. Many public datasets, including ChestX-ray14, use NLP to extract labels from free-text radiology reports, leading to incorrect or ambiguous label assignments [3].

Label co-occurrence complicates learning. Some conditions naturally occur together, such as “Atelectasis” and “Effusion,” which can confuse models trained independently per class [2].

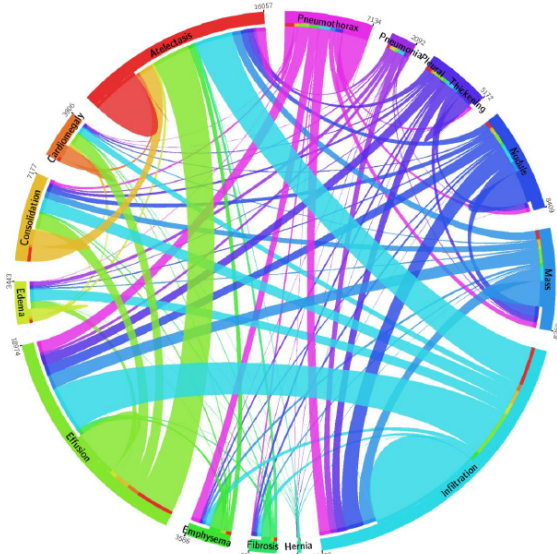


Figure 2.2. Disease label distribution in the ChestX-ray14 dataset. Class imbalance is a major concern in multi-label classification [1].

2.4 NIH ChestX-ray14 Dataset and Benchmark Studies

The ChestX-ray14 dataset, introduced by Wang et al., contains 112,120 frontal-view X-rays from over 30,000 patients. It is one of the most widely adopted benchmarks for multi-label CXR classification. Labels were

derived using automatic NLP on radiology reports, creating a weakly supervised but scalable labeling process [1].

Wang et al. also established initial benchmarks by training and evaluating several standard CNN architectures (e.g., AlexNet, VGG-16, GoogLeNet, and ResNet-50). The performance of these models was evaluated using AUC-ROC curves across all 14 disease classes.

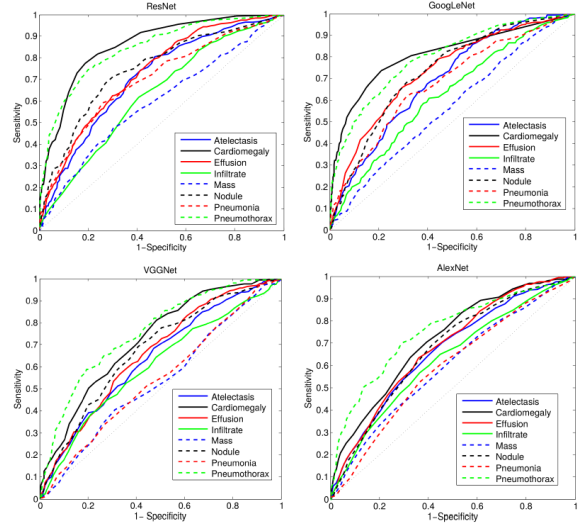


Figure 2.3. AUC-ROC performance of four CNN architectures on ChestX-ray14 [1].

2.5 Our Project in Context

Many high-performing models for CXR classification rely on large architectures like DenseNet121 or ensembles trained with auxiliary clinical inputs. These models, while powerful, are expensive to train and difficult to deploy in real-world clinical environments. Our project aims to improve on this by using MobileNetV2 as a lightweight yet competitive backbone [5].

To mitigate imbalance, we removed extreme cases like No Finding and Hernia, applied class-aware oversampling, and implemented a margin-based loss to improve label confidence calibration. We also applied dynamic thresholding, allowing the model to fine-tune decision boundaries per label rather than using a fixed threshold for all outputs.

3 Methodology

3.1 Model

The model uses MobileNetV2 as the backbone of the architecture. MobileNetV2 is lightweight in design, making it efficient to train [6]. This is ideal as we do not have access to a high-performance computing backend. It is also quicker to train than other models, enabling us to run more tests and iterate faster.

We used the default pretrained weights that PyTorch provides. These weights were trained on ImageNet, a large visual database with over 14 million images [7]. We used these weights because they provide the model with a solid foundation of identifying shapes, which transfer well to other domains such as chest X-ray classification, especially when domain-specific data is limited.

The first convolutional layer in the model was modified to accept grayscale input instead of RGB. X-ray images lack color, so eliminating the RGB channels helps reduce complexity and improves training efficiency.

We added fully connected layers with batch normalization and dropout (set to 0.5) to improve generalization and reduce overfitting. Only three dense layers were used, going from 1280 to 512 to 13 nodes (for 13 diseases), because the NIH dataset is small after removing the “No Finding” class [8].

3.2 Dataset and Preprocessing

We selected the NIH ChestX-ray14 dataset because it is a standard multi-label benchmark for evaluating medical imaging models [1]. The dataset contains over 100,000 frontal-view chest X-rays from more than 30,000 patients [8]. However, the dataset is highly imbalanced: over 60,000 X-rays (more than 60%) are labeled “No Finding.” Including this label would bias the model to favor that prediction. To address this,

we removed all “No Finding” samples. We also discarded the “Hernia” class, which has only 200 examples. It is too few to train reliably.

In addition to its popularity, we chose ChestX-ray14 because its moderate size allows for fast training, making experimentation more efficient. Integrating other datasets, like MIMIC-CXR [11], introduces labeling inconsistencies. Starting with a cleaner baseline allows us to evaluate methods more clearly before merging datasets.

All images were resized to 128×128 pixels to reduce computational cost while preserving relevant radiological features.

Data augmentation was applied using random horizontal flips, slight affine transformations (5° rotation, 2% scaling and translation), and minor brightness adjustments. This makes the model more robust to variations in patient positioning and image capture conditions.

3.3 Loss

The model was designed to produce high-confidence predictions for true labels and low-confidence predictions for false ones. To encourage this, we implemented a custom margin-based loss. Labels present in an image are encouraged to exceed a confidence threshold of 0.6, while absent labels are penalized if confidence exceeds 0.4. This reduces both false negatives and false positives by applying stronger penalties to uncertain or incorrect predictions.

3.4 Training Setup

We used the Adam optimizer for training, as it adapts the learning rate for each parameter, which speeds up convergence and improves performance [10]. Adam’s automatic step-size adjustment eliminates the need for manual learning rate tuning.

A learning rate of 0.0005 was selected to balance convergence speed and training sta-

bility. Higher values risk overshooting minima, while lower rates can lead to sluggish learning. Adam typically performs well with smaller learning rates [10], making 0.0005 a reasonable default.

We trained the model for 50 epochs with early stopping, halting when validation loss no longer improved. This prevents overfitting and ensures efficient resource usage.

3.5 Class Imbalance Handling

Most CXR datasets, including ChestX-ray14, exhibit strong class imbalance [1]. Conditions like Effusion are common, while others such as Pneumonia or Mass are much rarer [11, 10]. Without adjustment, models naturally bias toward frequent classes and neglect rare ones.

To mitigate this, we computed class weights using the ratio of negative to positive examples per class:

$$\text{Class Weight} = \frac{\text{Num of Negative Samples}}{\text{Num of Positive Samples}}$$

These weights were passed into the binary cross-entropy (BCE) loss function to emphasize rare classes during training. As a result, the model is penalized more strongly for misclassifying underrepresented conditions, which improves balanced performance.

4 Experimental Results

4.1 Overview of Experimental Design

The experiments followed a progressive tuning approach. We began with a baseline ResNet18 model without any pretraining or optimization to establish a reference point. Each subsequent experiment introduced one or more modifications, such as pretrained weights, custom loss functions, or data rebalancing techniques, with the goal of improving classification accuracy

and robustness. Our evaluation focused on three metrics: subset accuracy, recall, and AUC-ROC, which are further explained in Section 5. A summary of all model variants and their corresponding performance is provided in Figure 4.1.

Experiment	Model	Changes Introduced	Recall	Accuracy	AUC-ROC
Exp 1	ResNet18 (baseline)	No pretraining, default BCE, no thresholding, no balancing	0.10	0.03	0.60
Exp 2	ResNet18 (tuned)	Pretrained, margin+BCE loss, oversampling, thresholding, layer freezing	0.35	0.38	0.76
Exp 3	DenseNet121	Pretrained only, no balancing, no loss tuning	0.10	0.10	0.60
Exp 4	MobileNetV2 (pretrained)	Lightweight, no loss tuning or balancing	0.21	0.22	0.68
Exp 5	MobileNetV2 (no pretrain)	Random init, no tuning	0.12	0.14	0.55
Exp 6	MobileNetV2 Final	Pretrained, margin loss, dynamic thresholding, oversampling, grayscale input, bias init	0.43	0.50	0.80

Figure 4.1. Summary of performance for each experiment. The table shows key metrics like recall, accuracy, and AUC-ROC for each model variant.

4.2 Experiment 1: ResNet18 (Baseline, No Pretraining)

The first experiment used a ResNet18 architecture trained from scratch using the standard binary cross-entropy (BCE) loss. No pretraining, thresholding, or data balancing techniques were applied. This version served as the control to evaluate the model’s raw ability to learn from the highly imbalanced dataset [1]. Performance was weak across the board. The model achieved a recall of approximately 0.10, an accuracy of 0.03, and an AUC-ROC of 0.60. These results highlighted the limitations of training a deep model on a sparse multi-label dataset without any form of balancing or prior initialization.

4.3 Experiment 2: ResNet18 with Rebalancing, Margin Loss, and Threshold Tuning

For the second experiment, used in our Rev0 presentation, we introduced several optimizations to the baseline ResNet18 configuration. First, we applied over- and undersampling to reduce label imbalance.

Next, we implemented dynamic thresholding to replace the default 0.5 cutoff, optimizing per-class thresholds based on validation performance. Additionally, a margin-based auxiliary loss was incorporated alongside BCE, penalizing the model for low-confidence predictions on positive labels. These modifications yielded a significant improvement in model performance. The model’s recall increased to 0.35, accuracy rose to 0.38, and AUC-ROC reached 0.76. This experiment demonstrated that architectural changes alone were not enough, effective loss functions, threshold calibration, and data balancing played a major role in multi-label disease detection [14].

4.4 Experiment 3: DenseNet121 (Pretrained Only)

The third experiment tested the DenseNet121 architecture with pretrained ImageNet weights. This model configuration did not use any additional tuning, rebalancing, or threshold optimization. It was intended to measure the benefits of deeper architectures without specialized adjustments for the dataset. Despite its depth and pretrained initialization, DenseNet121 showed no improvement over the tuned ResNet18. The recall and accuracy remained at 0.10, with an AUC-ROC of 0.60. These results indicated that architectural complexity without targeted optimization strategies did not meaningfully impact performance on this task.

4.5 Experiments 4 and 5: MobileNetV2 with and without Pretraining

Given the need for efficient inference and faster training cycles, we next transitioned to the MobileNetV2 architecture. In Experiment 4, the model was initialized with pretrained ImageNet weights. In Experiment 5, we tested the same architecture without any pretraining, training it from scratch on the full dataset. The pretrained Mo-

bileNetV2 model achieved a recall of 0.21, accuracy of 0.22, and AUC-ROC of 0.68. In contrast, the non-pretrained version performed noticeably worse, with a recall of 0.12, accuracy of 0.14, and AUC-ROC of 0.55. These outcomes confirmed the importance of using pretrained weights, especially when dealing with sparse training signals from low-frequency labels [6].

4.6 Experiment 6: Final MobileNetV2 with Full Optimization

The final model, used in our Rev1 presentation, combined all validated enhancements. We used pretrained MobileNetV2 with grayscale input to reduce unnecessary complexity in early convolutional layers. Dynamic thresholding was applied for each class using validation data. The loss function combined binary cross-entropy with a margin-based penalty to encourage confident label predictions. Over- and undersampling techniques were used to rebalance the dataset, and we removed two problematic labels (“No Finding” and “Hernia”) that skewed training due to extreme imbalance. We also initialized the final layer’s bias using log-odds to reflect label prevalence. This fully optimized configuration delivered the highest overall performance. The model achieved a recall of 0.43, accuracy of 0.50, and an AUC-ROC of 0.80, making it the most successful and deployment-ready architecture of the experiments [15].

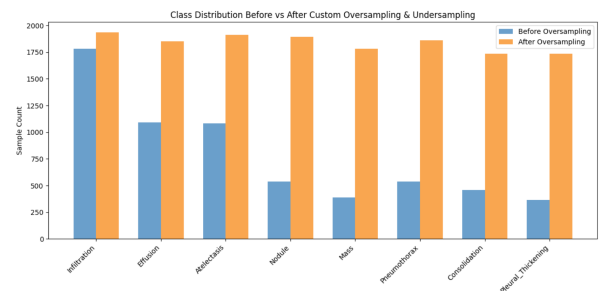


Figure 4.2. Histogram of label distribution before and after applying oversampling and undersampling.

Balancing the training set helped expose the model to underrepresented disease classes.

4.7 Impact of Dynamic Thresholding

In addition to improved architecture and balancing strategies, we found that replacing the default static threshold with dynamic, class-wise thresholds significantly improved model performance. This was particularly important in cases where class prevalence was low, as static thresholds tended to suppress true positives for rare conditions [14].

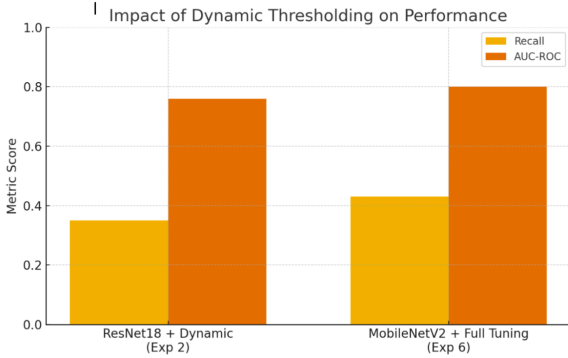


Figure 4.3. Bar plot showing Recall and AUC-ROC for the ResNet18 model (Experiment 2) and final MobileNetV2 model (Experiment 6), highlighting the improvement gained through dynamic threshold tuning.

4.8 Final Comparison Across Experiments

To provide a holistic view of our experimental outcomes, we compiled the final metrics for all six model variants. The progression of results illustrates how each layer of tuning, from pretrained weights to loss function, balancing, and threshold calibration, contributed to our final model’s performance. The comparison is illustrated in Figure 4.3.

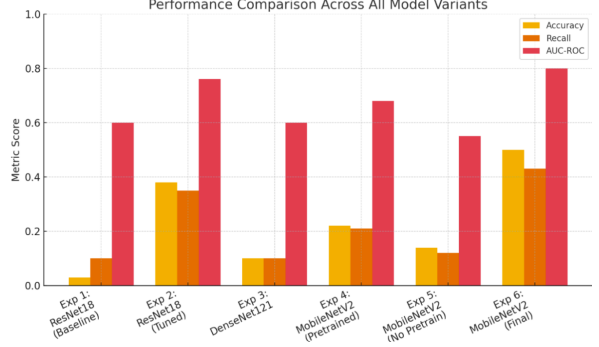


Figure 4.4. Grouped bar chart comparing accuracy, recall, and AUC-ROC across all six experiments. The final MobileNetV2 configuration consistently outperformed earlier models across all metrics.

5 Experimental Results

5.1 Performance Trends Across Experiments

Model performance improved steadily with each experimental refinement. The baseline ResNet18 model (Experiment 1), trained from scratch without pretraining or data balancing, exhibited poor performance on all metrics. Its limited recall and low AUC-ROC underscored the challenges of training deep models on sparse, multi-label datasets without guidance from prior weights or balancing mechanisms.

Experiment 2 introduced several critical enhancements, including data rebalancing, a margin-based auxiliary loss, and dynamic thresholding. These changes yielded substantial gains, with recall improving to 0.35 and AUC-ROC rising to 0.76, highlighting the significant impact of label-sensitive tuning [14].

DenseNet121 (Experiment 3), although deeper and initialized with pretrained ImageNet weights, failed to exceed the performance of the tuned ResNet18. This result reinforced that architectural depth alone is not a substitute for domain-appropriate adjustments.

Experiments 4 through 6 evaluated the Mo-

bileNetV2 architecture under varying levels of optimization. The final version (Experiment 6), which combined pretrained initialization, loss modification, threshold calibration, and grayscale input, achieved the strongest overall metrics: 0.43 recall, 0.50 accuracy, and 0.80 AUC-ROC.

Figure 4.4 previously showed these performance trends across all six experiments, emphasizing the effectiveness of lightweight architectures paired with tailored tuning strategies.

5.2 Effect of Threshold Optimization

Among all optimization techniques, threshold calibration had the most consistent and measurable impact. Early models relied on a uniform threshold of 0.5, which led to systematic under-detection of less prevalent pathologies. By dynamically adjusting thresholds for each label using validation data, recall improved without materially compromising classification confidence [14].

This effect is evident in Figure 4.3, where recall increased by 0.08 and AUC-ROC rose by 0.04 from Experiment 2 to Experiment 6. Dynamic thresholding enabled the model to adapt to varying class distributions, making it more robust in handling ambiguous or minority-class predictions, a crucial factor for deployment in diagnostic workflows.

5.3 Final Model Performance

The fully optimized MobileNetV2 model emerged as the top performer across all metrics. An AUC-ROC of 0.80 confirmed the model’s ability to reliably rank disease vs. non-disease cases. Meanwhile, a recall of 0.43 marked a significant gain in sensitivity relative to the baseline’s 0.10. These improvements, combined with MobileNetV2’s computational efficiency, suggest the model is well-suited for real-time clinical screening tools.

While this study focused on multi-label classification, the next logical step is to incorporate localization mechanisms. Visual explanations such as class activation maps (CAMs) and bounding boxes can improve interpretability and user trust in clinical contexts.

Figure 5.1 illustrates this future direction using an example from the ChestX-ray8 dataset [1]. It shows how diagnostic labels can be associated with regions in the image through weakly supervised localization, providing a benchmark for future explainable AI models in radiology.

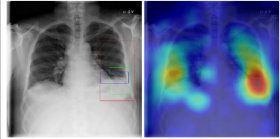
Radiology report	Keyword	Localization Result
findings include: 1. left basilar atelectasis/consolidation. 2. prominent hilum (mediastinal adenopathy). 3. left pic catheter (tip in atriocaval junction). 4. stable, normal appearing cardiomeastinal silhouette. impression: small right pleural effusion otherwise stable abnormal study including left basilar infiltrate/atelectasis, prominent hilum, and position of left pic catheter (tip atriocaval junction).	Effusion; Infiltration; Atelectasis	

Figure 5.1: Weakly-supervised localization in chest X-rays. The table shows the radiology report with keywords like “Effusion” and “Infiltration,” while the heatmap highlights the localized regions of predicted conditions on the X-ray, aiding in diagnostic interpretation.

Though our current system does not implement localization, this remains an essential milestone for future iterations aimed at enhancing radiologist support and transparency.

6 Analysis and Discussion

The experimental results presented in Section 5 reveal a clear trend: performance improvements were strongly correlated with the application of domain-specific strategies that addressed the inherent challenges of medical image classification. Among these, methods targeting class imbalance, decision thresholding, and prediction confidence had the most pronounced impact on recall and AUC-ROC—two metrics critical for clinical reliability.

6.1 Why Some Models Performed Better

The baseline ResNet18 model (Experiment 1) failed to generalize, primarily due to the lack of pretraining and the severe label imbalance present in the ChestX-ray14 dataset [1]. Its low recall and accuracy reflect a tendency to favor dominant classes while neglecting less frequent but clinically important conditions. This pattern aligns with previous studies that highlight the difficulty of training deep models on highly skewed medical datasets [3].

DenseNet121 (Experiment 3), despite its architectural depth and use of pretrained ImageNet weights, also underperformed. While such models are effective on natural image tasks, they often fall short in medical domains without targeted optimization. This observation is consistent with findings in the NIH and MIMIC-CXR literature, which suggest that deeper architectures alone are insufficient when faced with label noise and class imbalance [1], [4].

In contrast, the MobileNetV2 architecture consistently outperformed both ResNet and DenseNet variants when paired with appropriate training techniques. Pretraining provided an initial performance boost in Experiment 4, but more meaningful improvements emerged only after integrating margin-based loss, oversampling, dynamic thresholding, and log-odds-based bias initialization in Experiment 6.

The custom loss function enhanced sensitivity by penalizing uncertain predictions, encouraging higher confidence for true positives [15]. Meanwhile, dynamic thresholding helped adjust for uneven class distributions, improving detection of underrepresented conditions that were previously suppressed by a static threshold.

6.2 Limitations and Model Behavior

Despite outperforming previous configurations, the final model exhibited several notable limitations. While recall improved to 0.43, this still implies that over half of true disease instances remained undetected. This shortfall is partly attributed to the inherent label noise in the ChestX-ray14 dataset, which relies on mined radiology reports and lacks consistent ground truth [1].

Additionally, performance may be constrained by co-occurrence conflicts in multi-label settings. Certain pathologies, such as Cardiomegaly and Effusion, often present overlapping radiographic features, complicating classification. Such ambiguities, frequently noted in prior studies, remain an unresolved challenge for many automated detection pipelines.

While oversampling improved exposure to rare classes, it also introduced some risk of overfitting, particularly in low-prevalence categories. This effect was mitigated through regularization techniques such as dropout and data augmentation, but it remains an area for deeper investigation in future work.

6.3 Relationship to Prior Work

When compared to previous CNN-based approaches on the ChestX-ray14 dataset, such as CheXNet [3], our results are competitive despite using a significantly lighter architecture. CheXNet, which achieved AUCs in the 0.80-0.84 range using DenseNet121 and extensive data curation, was computationally intensive. In contrast, our MobileNetV2 model reached an AUC of 0.80 with fewer parameters and faster training, offering a more practical solution for deployment in resource-constrained or edge-based clinical settings.

Although recent research has begun exploring transformers and diffusion models for medical imaging, our own exploratory work

with diffusion-based approaches was limited by implementation complexity and computational cost. However, as toolsets for such methods become more accessible, they represent a promising direction for future iterations of this work.

7 Conclusion & Future Work

This project focused on developing and iteratively refining a deep learning system for multi-label classification of chest radiographs, using the NIH ChestX-ray14 dataset [1]. The aim was to produce a lightweight, accurate, and clinically meaningful model suitable for diagnostic support in real-world healthcare environments.

Starting with a baseline ResNet18 architecture, we progressively introduced a series of targeted enhancements. These included data rebalancing via over- and undersampling, label-wise threshold optimization, and a custom margin-based loss function to increase the model’s confidence in disease detection.

Through these experiments, the model demonstrated consistent gains across all evaluation metrics. The final version-based on pretrained MobileNetV2 with grayscale inputs, dynamic thresholds, and bias initialization-achieved an accuracy of 0.50, a recall of 0.43, and an AUC-ROC of 0.80. These results reflect strong generalization, particularly in the context of imbalanced and noisy training data.

Perhaps most notably, the results highlight that efficient and explainable architectures can still deliver competitive performance when paired with clinically informed train-

ing strategies. This has meaningful implications for diagnostic deployment, especially in low-resource settings where computational capacity is limited.

7.1 Future Work

While the model achieved promising results, there are several opportunities for continued development and refinement.

A critical next step is direct collaboration with radiologists. Expert input could support both interpretability and labeling quality, helping to define clearer diagnostic criteria, reduce noise in training data, and align model predictions with clinical expectations. It would also enable the design of more practical evaluation protocols, focusing on utility in real-world diagnostic workflows.

Expanding the training dataset with higher-quality or more diverse imaging sources, such as MIMIC-CXR [4] or hospital-based archives, would likely improve generalization, particularly in the detection of rare or visually overlapping pathologies.

Additionally, future work may explore more advanced architectures such as transformers or diffusion models. Although these approaches were not fully implemented in this study due to computational constraints, they hold potential for capturing more nuanced patterns in radiographic data, especially when supported by improved annotation quality.

Taken together, this work establishes a strong foundation for scalable, interpretable, and deployable AI systems in clinical imaging, while also identifying key research directions for future progress.

References

- [1] 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 *Proc. CVPR*, 2017, pp. 3462-3471. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2017/papers/Wang_ChestX-ray8_Hospital-Scale_Chest_CVPR_2017_paper.pdf
- [2] I. Baltruschat, H. Nickisch, M. Grass, T. Knopp, and A. Saalbach, “Comparison of deep learning approaches for multi-label chest X-ray classification,” *Sci. Rep.*, vol. 9, no. 1, p. 6381, 2019. Available: <https://www.nature.com/articles/s41598-019-42294-8>
- [3] J. Irvin *et al.*, “CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison,” in *Proc. AAAI*, vol. 33, no. 1, 2019, pp. 590-597.
- [4] A. E. Johnson *et al.*, “MIMIC-CXR: A large publicly available database of labeled chest radiographs,” *arXiv preprint arXiv:1901.07042*, 2019. [Online]. Available: <https://physionet.org/content/mimic-cxr/2.1.0/>
- [5] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” in *Proc. CVPR*, 2018, pp. 4510-4520. [Online]. Available: <https://arxiv.org/abs/1801.04381>
- [6] A. Howard *et al.*, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications” *arXiv preprint arXiv:1905.02244*, 2019. [Online]. Available: <https://arxiv.org/abs/1704.04861>
- [7] “ImageNet: A large-scale hierarchical image database.” [Online]. Available: <https://www.image-net.org/about.php>
- [8] “NIH Clinical Center provides one of the largest publicly available chest X-ray datasets to the scientific community,” National Institutes of Health. [Online]. Available: <https://www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available-chest-x-ray-datasets>
- [9] P. Rajpurkar *et al.*, “CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning” Stanford University, 2017. [Online]. Available: <https://stanfordmlgroup.github.io/projects/chexnet/>
- [10] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization” in *arXiv*, 2015. [Online]. Available: <https://arxiv.org/abs/1412.6980>
- [11] Johnson, A., Pollard, T., Mark, R., Berkowitz, S., & Horng, S. (2024). MIMIC-CXR Database (version 2.1.0). PhysioNet. <https://doi.org/10.13026/4jqj-jw95>
- [12] Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J. et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data* 6, 317 (2019). <https://doi.org/10.1038/s41597-019-0322-0>
- [13] Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P. C., Mark, R., Stanley, H. E., et al. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* [Online]. 101 (23), pp. e215–e220.

- [14] I. Pillai, G. Fumera, F. Roli, “Threshold optimisation for multi-label classifiers” *ScienceDirect*, July 2013. [Online]. Available: <http://dx.doi.org/10.1016/j.patcog.2013.01.012>
- [15] H. Zhou, *et al.*, “Comparing to Learn: Surpassing ImageNet Pretraining on Radiographs by Comparing Image Representations” in *Proc. MICCAI*, 2020, pp. 398-407. [Online]. Available: https://doi.org/10.1007/978-3-030-59710-8_39