PROJECT FINAL REPORT

Adam Roberge
Student# 1009491186
adam.roberge@mail.utoronto.ca

Bill Jiang
Student# 1008998907
bill.jiang@mail.utoronto.ca

Chris (Seoyoung) Kwon
Student# 1005543789
christine.kwon@mail.utoronto.ca

Mitchell Souliere-Lamb
Student# 1007060305
mitchell.souliere.lamb@mail.utoronto.ca

1 INTRODUCTION

1.1 MOTIVATIONS BEHIND THE PROJECT

In recent years, the field of medical imaging has seen significant advancements through the application of machine learning, particularly deep learning techniques (Gao et al. (2023))(Zhang et al. (2023))(Liu et al. (2023)). This project focuses on developing a model to classify chest X-ray images into various categories, such as different types of lung diseases. The motivation behind this project lies in the potential for machine learning to assist radiologists in diagnosing diseases more accurately and efficiently, especially in regions with a shortage of medical professionals.

1.2 SINCE THE PROJECT PROGRESS REPORT

In the initial stages of this project, we utilized a smaller dataset consisting of 4,000 endoscopy images and achieved over 95% accuracy using a U-Net-inspired classification model (Ronneberger et al. (2015))(Mei (2024)) that we developed. Building on this success, we decided to take the project further by tackling a more challenging problem. We expanded our scope to include a larger and more complex dataset of chest X-ray images. To address the increased data complexity, we also enhanced our model's architecture to improve its performance on this more advanced task. All codes are present on ourGitHub repository

1.3 PROJECT GOALS

The primary objectives of this project are outlined below:

- Develop a deep learning model to classify chest X-ray images into various categories, including different lung diseases.
- Improve diagnostic accuracy and efficiency in identifying lung diseases.
- Provide a tool to assist radiologists in making quicker and more accurate diagnoses.

2 BACKGROUND & RELATED WORKS

Deep learning has significantly advanced medical imaging, with CNNs like UNet being fundamental in image segmentation and classification. Introduced by Ronneberger et al., UNet's encoder-decoder structure is adept at capturing both high- and low-level features, making it a staple in medical diagnostics(Ronneberger et al. (2015)).

Recent advancements include transformer-based architectures and self-supervised learning. Gao et al. demonstrated that transformers excel at capturing long-range dependencies, crucial for analyzing complex medical images like chest X-rays (Gao et al. (2023)). Liu et al. further highlighted transformers' potential to enhance diagnostic accuracy in lung disease classification (Liu et al. (2023)). These techniques often surpass traditional CNNs in specific tasks.

Self-supervised learning has also emerged as a key advancement, enabling models to learn from unlabeled data, which is particularly useful in medical imaging where labeled data is scarce. Zhang et al. showed that this approach reduces the need for large labeled datasets (Zhang et al. (2023)). Shao and Wang found that combining self-supervised learning with enhanced softmax functions can improve classification performance (Shao & Wang (2023)).

However, the application of attention mechanisms in medical imaging has produced mixed results. Uddin's attention-based DenseNet suggested potential benefits (Uddin (2024)), but our experiments showed that attention layers added complexity without performance gains, aligning with findings by Yi et al. (Yi et al. (2020)). Therefore, we did not incorporate attention mechanisms in our final model.

This project builds on UNet's foundation, leveraging scaling techniques from the EfficientNet (Tan & Le (2019)) to enhance chest X-ray classification while acknowledging the limitations of attention mechanisms in this context.

3 Data Preprocessing & Processing

3.1 Initial Stages

We initially attempted to merge the NIH Chest X-ray (112,000 images) and CheXpert (224,000 images) datasets but faced significant issues with mislabeled entries, missing data, and inconsistencies in the .csv files. Due to these challenges, we focused solely on the NIH dataset, further filtering it to approximately 40,000 images by retaining only those with a single disease label.

3.2 Data Imbalance of the NIH Chest X-ray Dataset and Data Cleaning

The NIH dataset exhibited significant class imbalance and included images with multiple disease labels [Figure 1], complicating the classification task. To streamline the process and focus on single-label classification, we filtered the dataset to retain only images associated with a single disease label. This approach simplified the model's task, enabling it to target the classification of one condition at a time and reducing prediction ambiguity.

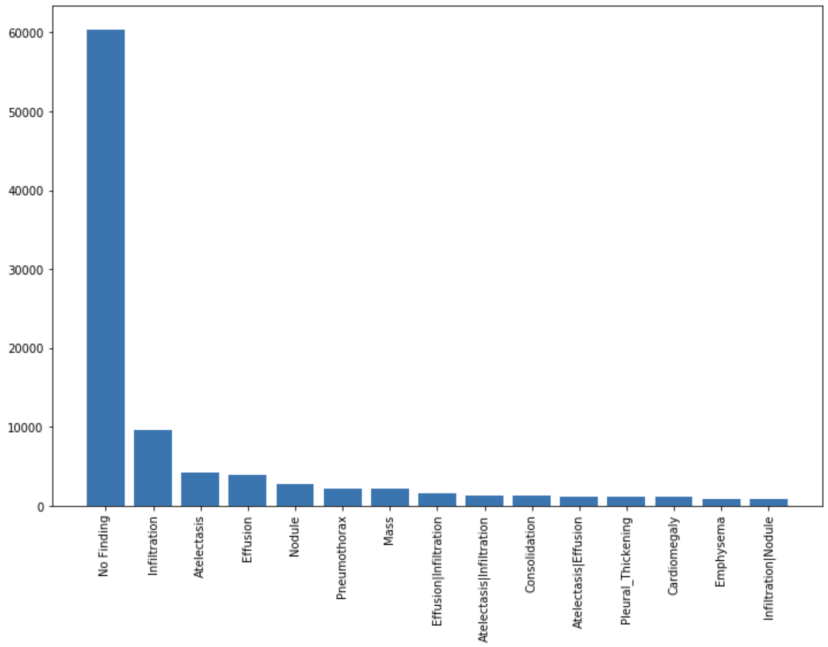

Figure 1: Data Imbalance of NIH Chest X-ray image data

To address the issue of data imbalance after filtering, we calculated the average number of images remaining per class and used this as a target to balance all classes. Specifically, we found that the average number of images across the remaining classes was 12,159. To achieve this balance, we applied the following steps:

1. **Class Filtering:** Classes with fewer than 2,000 images were excluded to ensure adequate sample sizes for effective training.
2. **Undersampling and Oversampling:** The "No Finding" class was reduced in size, and other classes were increased to match the average of 12,159 images per class, involving the addition or removal of thousands of samples to achieve balance.
3. **Final Dataset:** This process resulted in a balanced dataset of 85,113 X-ray images, with 7 classes [Figure 2], each containing exactly 12,159 images [Table 1]. A corresponding .csv file was generated with image titles and their respective classes.

The table below presents the number of samples that were added or removed per class during this process:

Class	Original Number of Images	Final Number of Images
Atelectasis	5,389	12,159 (+6,770)
Effusion	4,763	12,159 (+7,396)
Infiltration	9,627	12,159 (+2,468)
Mass	2,842	12,159 (+9,317)
No Finding	54,379	12,159 (-42,220)
Nodule	3,687	12,159 (+8,472)
Pneumothorax	2,942	12,159 (+9,217)

Table 1: Image Number Adjustments for Each Class

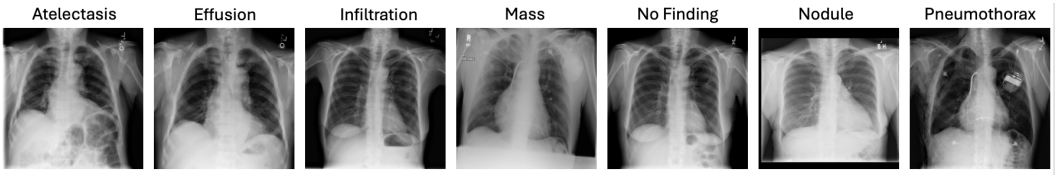

Figure 2: Sample images from each class used in this project from the NIH Chest X-ray dataset

This approach allowed us to create a balanced and manageable dataset for single-label classification, which was essential for training our deep learning model effectively and ensuring reliable, unbiased classification performance.

3.3 Label Correction and CSV File Generation

To ensure dataset accuracy, we generated a new .csv file reflecting the cleaned and balanced data. This involved reorganizing images into class-specific folders and updating the .csv file with accurate, single-label classifications, providing a reliable reference for model training.

3.4 Class Mapping and Dataset Creation

We standardized label representation by mapping class names to unique integer indices, ensuring consistency in model training and evaluation. This mapping was applied during both the training and testing phases.

3.5 Augmentation, Data Splitting, and Data Loading

To enhance generalization, we applied data augmentation techniques such as resizing, cropping, flipping, rotation, color jittering, affine transformations, and random erasing during training. The

dataset was split into training, validation, and test sets (70/15/15 ratio) with stratification to maintain class balance. Separate .csv files were created, and custom PyTorch datasets were configured for efficient batch processing. The test set remained separate to accurately measure the model's performance on unseen data.

- **Training Data:** Loaded *with* data augmentation to enhance model generalization.
- **Validation Data:** Loaded *without* data augmentation, used for hyperparameter tuning and model validation.
- **Testing Data:** Loaded *without* data augmentation, only resized and normalized, to provide an unbiased evaluation of the final model's performance

This systematic approach to data preprocessing and processing was critical in preparing the dataset for training the deep learning model, ensuring that the model would be exposed to diverse and balanced data, and that the training process would be as efficient and effective as possible.

4 ARCHITECTURE

4.1 APPROACHES TO ENHANCE THE MULTILEVELUNET

To optimize our initial MultiLevelUNet architecture, we explored two main enhancements: incorporating attention layers inspired by the Vision Transformer (ViT) and integrating scaling techniques from EfficientNet.

4.2 EXPLORATION OF ATTENTION LAYERS

We experimented with attention layers to enhance the model's focus on critical regions of chest X-ray images (Vaswani et al. (2017)). However, this approach increased training time significantly without improving performance, likely due to the localized nature of chest X-ray diseases (Chen et al. (2019)). Consequently, we excluded attention layers from the final model.

4.3 OPTIMIZING THE MULTILEVELUNET WITH SCALING TECHNIQUES

Recognizing the limitations of attention mechanisms, we focused on optimizing the MultiLevelUNet using EfficientNet-inspired scaling techniques (Tan & Le (2019)). This involved systematic scaling of the network's width, depth, and resolution to enhance capacity and efficiency:

- **Width Scaling**: Increased the number of convolutional filters by 1.2x, allowing the network to capture more features concurrently.
- **Depth Scaling**: Doubled the number of residual blocks, expanding the architecture from 35 to 71 layers, enhancing hierarchical learning.
- **Resolution Scaling**: Adjusted input resolution by 1.2x, balancing detail capture with computational efficiency.
- **Compound Scaling**: Applied compound scaling to optimize the balance between these three dimensions (Tan & Le (2021)).

4.4 MODEL ARCHITECTURE: EFFICIENT MULTILEVELUNET

The Efficient MultiLevelUNet builds upon the U-Net structure (Ronneberger et al. (2015)), incorporating 71 layers, skip connections, and EfficientNet-inspired(Tan & Le (2019)) scaling techniques [Figure 3]. This architecture is designed to meet the challenges of medical image classification, particularly chest X-rays.

4.4.1 CORE COMPONENTS

Double Convolution Block: The basic building block of the network is the DoubleConv module, which is derived directly from the original U-Net architecture (Ronneberger et al. (2015)). This

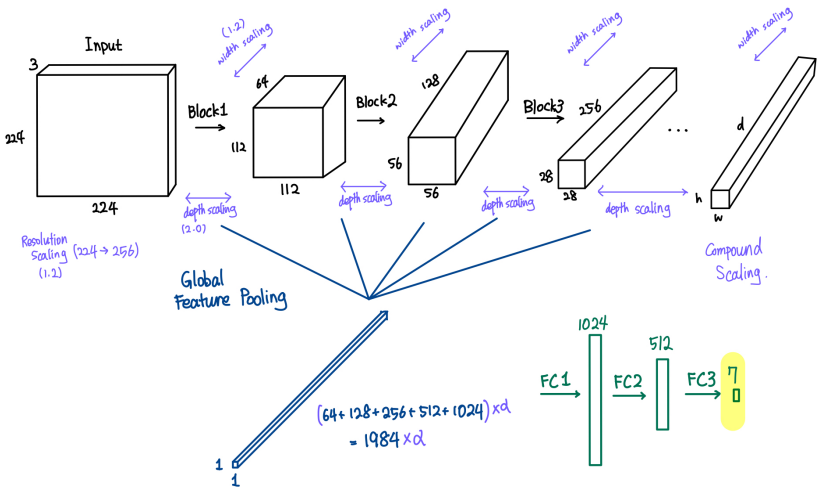

Figure 3: Original MultiLevelUNet Architecture (used in the progress report) with EfficientNet-inspired modifications highlighted in purple

module consists of two convolutional layers, each followed by batch normalization and ReLU activation, and remains unchanged in our adaptation. It enhances feature extraction at every level of the network, providing a robust foundation for more advanced layers.

Residual Block: To improve learning and convergence, we incorporated Residual Blocks with skip connections (He et al. (2016)), which help mitigate the vanishing gradient problem and preserve essential features as the network depth increases.

Downsampling Path: The downsampling path uses multiple residual blocks with max-pooling layers, progressively reducing spatial dimensions while increasing feature map depth. Feature maps from each level are stored for later concatenation.

Bottleneck: The bottleneck layer employs an expanded residual block, doubling the filters from the previous level, capturing the most abstract features before upsampling.

Feature Concatenation and Global Pooling: The network concatenates all feature maps obtained from the downsampling path and the bottleneck. These concatenated features undergo global average pooling, reducing each feature map to a single value, which allows the model to aggregate global context information across all abstraction levels.

Fully Connected Layers: The pooled features are passed through a series of fully connected layers, where the number of neurons is scaled according to the width coefficient. This section is responsible for the final classification task, translating the extracted features into class probabilities.

4.4.2 Summary

The Enhanced MultiLevelUNet integrates 71 layers with skip connections and EfficientNet-inspired scaling (Ronneberger et al. (2015); Tan & Le (2019)), optimizing width, depth, and resolution. This balance of computational efficiency and pattern recognition makes it well-suited for chest X-ray classification, with architectural enhancements like residual blocks and feature concatenation improving diagnostic performance.

5 Baseline Model: Pre-trained ResNet18

For this project, we employed a pre-trained ResNet18 model as the baseline to benchmark the performance of our Enhanced MultiLevelUNet architecture. ResNet18, known for its residual blocks that mitigate the vanishing gradient problem, was chosen due to its proven effectiveness in deep

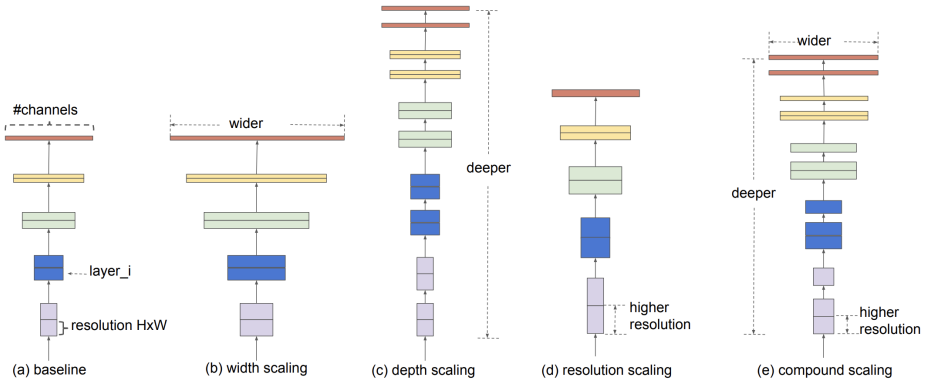

Figure 4: EfficientNet (Tan & Le (2019)

network training (He et al. (2016)). Pre-trained on the ImageNet dataset, ResNet18 offers a strong foundation of features applicable to various image types, including medical images.

We fine-tuned the final fully connected layer to align with the seven classes in our chest X-ray dataset, making ResNet18 a robust baseline. Comparing it against our custom architecture allowed us to measure the performance gains from the width, depth, and resolution scaling incorporated into the MultiLevelUNet model.

6 Quantitative Results

The performance of our models was evaluated using accuracy and loss as the primary quantitative measures, both of which were tracked across training and validation phases.

6.1 Efficient MultiLevelUNet Results

For the Efficient MultiLevelUNet model, we used a batch size of 32, a learning rate of 1×10^{-4}, and trained the model for 100 epochs with 4 data loader workers. Early stopping was implemented with a patience of 5 epochs and a minimum delta of 0.0005, which helped prevent overfitting by halting training when the validation loss did not improve sufficiently.

The Efficient MultiLevelUNet model showed consistent improvements in key performance metrics throughout training [Figure 5]: :

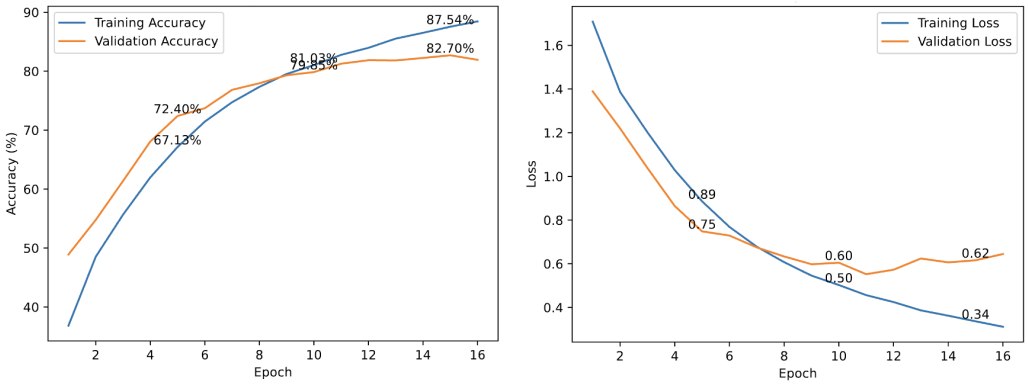

Figure 5: Efficient MultiLevelUNet Training and Validation Accuracy and Loss

- **Training Accuracy**: The model's accuracy steadily increased, reaching 87.54% by the final epoch, demonstrating strong learning capability.
- **Validation Accuracy**: The validation accuracy peaked at 82.70%, indicating effective generalization to unseen data.
- **Training Loss**: Loss consistently decreased from 1.6 to 0.34, reflecting effective error minimization.
- **Validation Loss**: Validation loss also decreased to 0.62, maintaining a balance with training metrics, suggesting minimal overfitting.

6.2 ResNet18 Baseline Results

The baseline ResNet18 model was trained with a batch size of 32, a learning rate of 1×10^{-5}, and also for 100 epochs using 4 data loader workers. The same early stopping criteria were applied, ensuring that the training was halted when further improvements in validation loss became negligible.

The ResNet18 baseline model also showed steady improvements [Figure 6]:

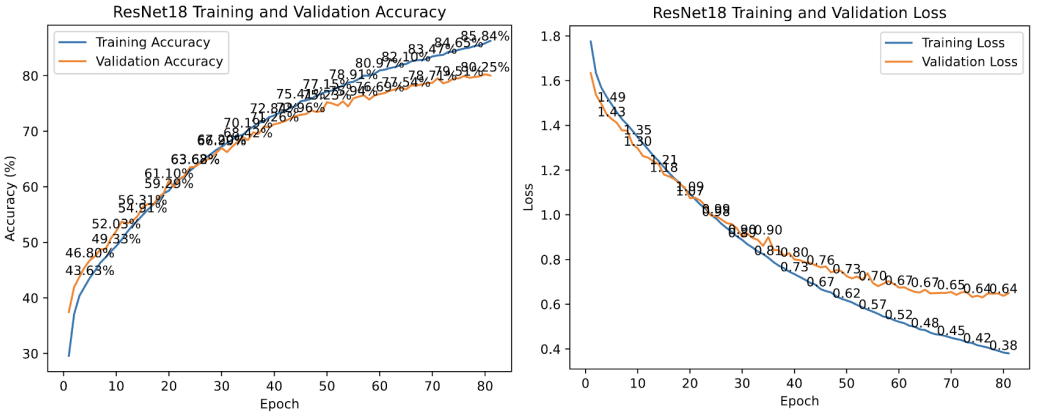

Figure 6: ResNet18 Training and Validation Accuracy and Loss

- **Training Accuracy**: Accuracy increased consistently, reaching 85.84% by the final epoch.
- **Validation Accuracy**: The validation accuracy reached 80.25%, slightly lower than the primary model but still indicative of good generalization.
- **Training Loss**: Loss decreased significantly to 0.38, showing the model's learning efficiency.
- **Validation Loss**: Validation loss plateaued at 0.64, indicating stable convergence.

6.3 Comparison and Insights

Overall, Both models performed strongly, with the Efficient MultiLevelUNet slightly outperforming the ResNet18 baseline in accuracy and loss. Notably, the Efficient MultiLevelUNet converged in just 16 epochs, compared to 81 for ResNet18, likely due to its advanced scaling techniques. These results confirm the enhanced architecture's effectiveness and efficiency, as well as the models' ability to generalize well to new data.

7 Qualitative Results

In addition to the quantitative metrics, qualitative analysis through confusion matrices and ROC curves further clarifies the performance of the models. The confusion matrices and ROC curves were generated based on the test set, which was completely isolated from the training and validation

processes. This ensures that the qualitative analysis reflects the model's performance on new, unseen data, providing an accurate assessment of its generalization capabilities

7.1 Efficient MultiLevelUNet Results

The qualitative performance of the Efficient MultiLevelUNet is depicted in the Confusion Matrix the Receiver Operating Characteristic Curve [Figure 7].

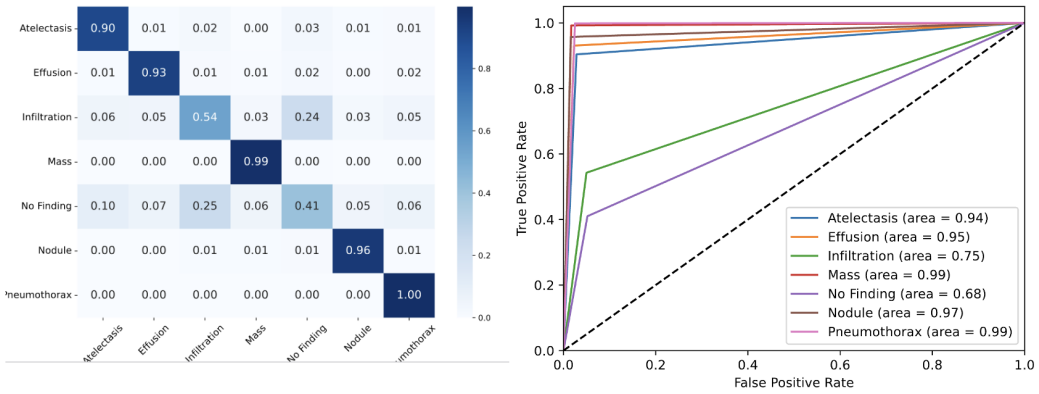

Figure 7: Efficient MultiLevelUNet Confusion Matrix and ROC Curve

- **Confusion Matrix**: The confusion matrix shows that the Efficient MultiLevelUNet model excels in accurately classifying the majority of the classes, particularly 'Mass' and 'Pneumothorax,' where it achieves almost perfect classification with minimal misclassifications. However, it struggles slightly with the 'Infiltration' class, which shows more confusion with other similar conditions like 'No Finding'.
- **ROC Curve**: The ROC curve indicates strong discriminative performance across all classes, with AUC values close to or exceeding 0.95 for most classes, except 'Infiltration' which has a lower AUC of 0.75. This suggests that while the model is highly effective overall, certain conditions that present with subtle variations in X-rays remain challenging.

7.2 ResNet18 Baseline Results

The ResNet18 baseline model's qualitative results are shown in the figure below [Figure 8]:

- **Confusion Matrix**: The confusion matrix reveals similar patterns, with high accuracy in identifying 'Mass' and 'Pneumothorax'. However, there is noticeable confusion between 'Infiltration' and other conditions, as well as 'No Finding,' indicating potential areas where the model could improve.
- **ROC Curve**: The ROC curve shows strong performance with AUC values above 0.90 for most classes. Similar to the Efficient MultiLevelUNet, 'Infiltration' is the most challenging class with a lower AUC, which points to the inherent difficulty in distinguishing this condition from others based on X-ray images.

7.3 Comparison and Insights

Overall, the qualitative results align with the quantitative findings, highlighting the Efficient MultiLevelUNet's superior performance across most classes, while also underscoring the challenges associated with classifying conditions like 'Infiltration'. These insights demonstrate the model's capability in generalizing well to various chest X-ray conditions, though it also points to areas for potential refinement in future iterations.

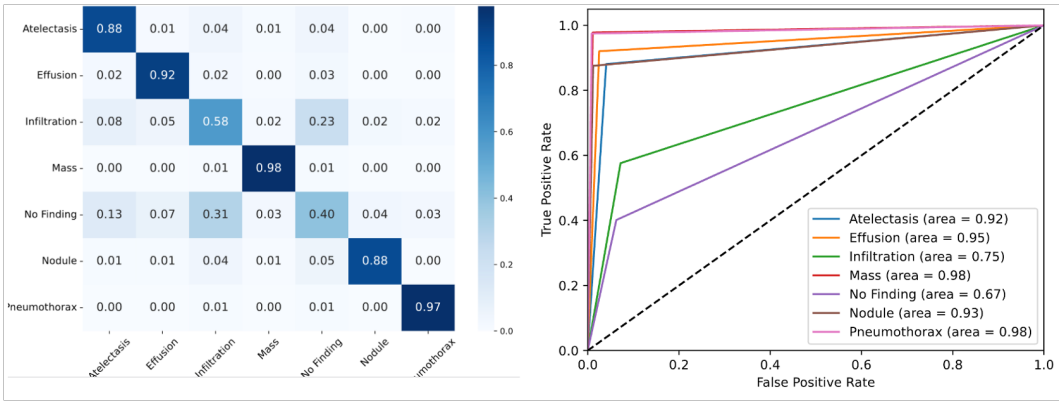

Figure 8: ResNet18 Confusion Matrix and ROC Curve

8 Discussion

The Efficient MultiLevelUNet model demonstrated strong performance in classifying chest X-ray images across multiple disease categories. High accuracy and consistent reductions in loss on both training and validation sets indicate effective learning of key disease patterns. The ROC curves and confusion matrices reinforce this, showing high true positive rates, particularly for critical conditions like Pneumothorax and Mass, where the model significantly outperformed the baseline.

A notable observation is the lower true positive rate for the Infiltration and No Findings classes. However, this outcome is positive in medical contexts. Infiltration, being less deadly, is less critical to detect with high precision, and occasional false positives in the No Findings class are preferable to missing a serious condition. This cautious approach aligns with medical diagnostics, where the cost of false negatives is much higher than false positives.

The model maintained a strong balance between precision and recall across most classes, crucial for reliable disease flagging. Overall, we are satisfied with the model's performance, particularly its ability to generalize to new data and its conservative bias, which is desirable in minimizing missed diagnoses in medical applications. The project underscored the importance of context in interpreting model results, with the model's cautious bias proving beneficial in a medical setting.

9 Ethical Considerations

The utilization of a machine learning model for medical image classification raises significant ethical considerations, particularly the potential for misdiagnosis. While our model performs well, it is not infallible. False positives could lead to unnecessary anxiety and treatment, while false negatives might result in missed diagnoses, which could be life-threatening. Additionally, the training data itself may contain inherent biases, such as underrepresentation of certain populations, which could lead to biased outcomes in real-world applications. Therefore, it is crucial that this model be used as an assistive tool for radiologists, rather than as a standalone diagnostic tool.

10 Project Difficulty / Quality

The most significant challenge in this project was preparing the dataset for effective training. Cleaning the data, creating a new balanced dataset, and performing both upsampling and downsampling to address class imbalances were time-consuming and required careful attention to detail. Additionally, generating a new, accurate .csv file to ensure correct label assignments was a critical and challenging step. Despite these obstacles, we successfully implemented a robust training process that led to high accuracy and strong generalization. Our model's performance exceeded expectations, demonstrating our ability to tackle complex tasks that extend beyond standard coursework in deep learning.

REFERENCES

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 932–944, 2019.

Jing Gao, Xiaohong Xie, and Yun Peng. Advances in deep learning for medical image analysis. *IEEE Transactions on Medical Imaging*, 42(4):1100–1113, 2023. doi: 10.1109/TMI.2023.3050124.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Cheng Liu, Yan Wang, and Feng Zhao. Transformers in medical imaging: A comprehensive review. *Computers in Biology and Medicine*, 152:106254, 2023. doi: 10.1016/j.compbiomed.2023.106254.

Song Mei. U-nets as belief propagation: Efficient classification, denoising, and diffusion in generative hierarchical models. *arXiv preprint arXiv:2404.12345*, Apr 2024. URL `https://arxiv.org/abs/2004.08790`. Submitted on 29 Apr 2024 (v1), last revised 1 May 2024 (this version, v2).

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *arXiv preprint arXiv:1505.04597*, May 2015. URL `https://arxiv.org/abs/1505.04597`. Submitted on 18 May 2015.

Hao Shao and Shunfang Wang. Deep classification with linearity-enhanced logits to softmax function. *Entropy*, 25(5):727, 2023. doi: 10.3390/e25050727. URL `https://doi.org/10.3390/e25050727`.

Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. *Proceedings of the International Conference on Machine Learning*, 2021.

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 6105–6114, 2019.

Jia Uddin. Attention-based densenet for lung cancer classification using ct scan and histopathological images. *Designs*, 8(2):27, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Xiaohua Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *IEEE Access*, 8:20453–20465, 2020. doi: 10.1109/ACCESS.2018.2817588.

Yang Zhang, Ming Li, and Jun Yan. Self-supervised learning in medical imaging: State-of-the-art and future perspectives. *Pattern Recognition Letters*, 167:50–60, 2023. doi: 10.1016/j.patrec.2023.01.014.