A study into the usage of Continual Learning in Thoracic Disease Classification

Mateus Becklas and **Megan van Drunen** and **Daniel Marin**

Damian Michailov and **Thorvald Rovers** and **Lonneke Zielinski**

Abstract

Thoracic diseases such as pneumothorax and nodules can be life-threatening if undetected, yet chest X-ray interpretation is complex and prone to human error. This study develops a Convolutional Neural Network using continual learning with Elastic Weight Consolidation (EWC) to classify chest X-rays into five categories: Atelectasis, Effusion, Infiltration, Nodule, and Pneumothorax. Using image upscaling and per-disease parameter fine tuning introduces statistically significant reductions in false negative rates across multiple pathologies. Emphasizing recall over raw accuracy addresses class imbalance and reduces harmful false negatives, aligning with the ethical principal of non-maleficence. Most promising configuration achieving recall above 0.75 on average across all five diseases with an average false negative rate below 0.25 demonstrating that responsible medical AI requires evaluating all design choices based on their impact, not just metrics.

1. Introduction

Thoracic disorders are diseases affecting the chest cavity. Early and accurate diagnosis is essential, as some conditions such as pneumothorax can become life-threatening without timely treatment, while others, such as lung nodules, require monitoring to prevent disease progression. Chest X-ray imaging is one of the most widely used diagnostic tools for detecting thoracic diseases because it is fast, cost-effective, and widely available. However, interpreting chest X-rays is complex and requires expert knowledge.

Recent studies show that deep learning models perform well in automated chest X-ray analysis. In particular, convolutional neural networks (CNNs) have demonstrated high performance in detecting patterns associated with thoracic diseases. For example, a study combining a ResNet-50 CNN with a Vision Transformer (ViT-B16) achieved an accuracy of 96.18% in multi-class classification of chest X-ray images for tuberculosis, viral pneumonia, and bacterial pneumonia (Hadhoud et al., 2024). These results suggest that deep learning models can assist clinicians in identifying diseases from medical images.

Considering these findings, this project develops a deep learning model to automatically classify chest X-rays into one of five categories: Atelectasis, Effusion, Infiltration, Nodule, or Pneumothorax. The project addresses technical challenges in model development while considering the ethical risks of diagnostic errors.

1.1 Problem description

The dataset used in this project is derived from the National Institutes of Health (NIH) Chest X-ray dataset and contains over 25,000 labeled and preprocessed X-ray images. The labels were automatically extracted from radiology reports using natural language process-

ing, which introduces potential label noise. Additionally, the dataset is highly imbalanced, with some diseases occurring much more frequently than others.

To address these challenges, a ResNet-inspired CNN is trained to classify thoracic disorders from chest X-ray images. Class imbalance is addressed using weighting and sampling strategies. The model is trained using a continual learning approach, where diseases are learned sequentially, while Elastic Weight Consolidation (EWC) prevents catastrophic forgetting of previously learned tasks.

Beyond technical challenges, the use of AI in medical diagnosis raises important ethical considerations such as accountability, reliability, fairness, and transparency. This project primarily focuses on the ethical principle of non-maleficence, which emphasizes minimizing harm to patients. Diagnostic models can produce false positives and false negatives, both of which may negatively affect patient outcomes.

1.2 Research questions

This study aims to investigate both the technical performance and the ethical implications of automated thoracic disease classification using deep learning. The main research question is: *How can a CNN model using continual learning classify thoracic diseases on chest X-rays be improved while minimizing harmful misdiagnosis (non-maleficence) for patients?* To address this question, the following sub-questions are explored:

1. How can the performance of a continual learning CNN for thoracic disease classification be evaluated to reflect the risk of harmful misdiagnosis?

2. How do dataset characteristics such as class imbalance and label noise affect the likelihood of harmful misdiagnosis?

3. Why is minimizing false negatives essential for AI-based diagnostic models in order to uphold the principle of non-maleficence?

2. Methodology and Experiments

2.1 Dataset analysis

The dataset for this project contains 25,262 X-ray images (128 x 128 pixels): 16,841 for training, 5,614 for testing, and 2,807 for final evaluation. Each image is labeled with one of six classes: five thoracic pathologies or "No Findings". This dataset presents two qualities which make the technical and ethical aspect more difficult. First, it is highly imbalanced, dominated by the "No Findings" class, leaving rarer conditions like nodule underrepresented (Figure 4). Ignoring this leads to high accuracy but many false negatives, violating non-maleficence by potentially harming patients. Secondly, the labels were mined using natural language processing, with an estimated accuracy of over 90%. This introduces label noise that limits model performance. These dataset characteristics (class imbalance and label noise) can increase the risk of harmful misdiagnoses, directly addressing our second sub research question.

2.2 Preprocessing

As a preprocessing step, the scikit-image library was used to upscale all the images from their original 128 x 128 size to a 256 x 256 resolution. The resize function performs spline interpolation to resize the N dimensional images, with each pixel being generated as a weighted average of surrounding input values. The upsampling was performed with the parameter preserve_range set to True to refrain the original pixel density without the use of normalization. We have identified the base 128 x 128 resolution as a limiting factor, particularly for pathologies such as nodule, where more details are critical for detection. By upscaling to 256 x 256, the model has more spatial information and the effect of this change is evaluated against the unprocessed data which operates at the original resolution.

2.3 Baseline Model

The provided baseline is a three-block convolutional neural network with a 6-way softmax output, mapping each input image to one of the six classes. And although the base architecture does make use of Batch Normalization, ReLU activations, and progressive dropout, there are still issues that limit this model for clinical utility. First of the issues is the successive max-pooling operations aggressively discard the spatial resolution, which is damaging for pathologies like nodules which are more localized and subtle. After that, the absence of non-linearity between two fully connected layers reduces the capacity for representation. Third, and most importantly, the softmax formulation enforces that the 6 classes are mutually exclusive, assigning the full probability mass to a single label. This is unrealistic under clinical use as multiple thoracic conditions can happen at the same time, it also biases the model towards the majority class under the imbalanced training data.

The results confirm this, accuracy is fairly high for the base model with four out of five pathologies achieving higher than 70% accuracy, yet recall is only at 25% and false negative rates are also high, more than 70%. This is majority class collapse, the model does not actually predict any cases of the pathologies despite appearing accurate. In connection with our perspective of sub research question 1, illustrating exactly why raw accuracy is an inappropriate metric in this setting. It hides a false negative rate that would not be acceptable in any clinical context and creates a false sense of security.

2.4 Experimental Comparison

To decide whether using a Continual Learning approach was appropriate, it was decided to do a comparison between the baseline model, a continual learning approach, and a continual learning approach with disease-specific parameters. In this capacity, an automation was first used to find the best performing parameters for the continual learning approach, and the disease-specific parameters. This automation chose its output based on a score given by the formula shown in Equation (1).

$$score = 0.60 \cdot F_1 + 0.30 \cdot Acc - 0.10 \cdot (FPR - FNR) - 0.40 \cdot degenerate\_task\_fraction \quad (1)$$

2.5 Experimental Results

We first look at the heatmap of the three models. In Figure 3 we can observe that the base model performs poor on Recall and F1, as they are predominantly orange/red colored.

Even though the average accuracy is relatively high, all other metrics perform poorly. This can also be observed in Figure 1. We then look at the heatmap in Figure 3 of the Continual Learning approach without disease-specific parameters. Here, it can be noted that the overall performance has increased. With the false negative rate dropping. Finally, we observe in the same heatmap that the Continual Learning Approach with disease-specific parameters performs more balanced. With the average false negatives staying under our threshold of 0.25, and the variance between diseases in FNR shrinking. From Figure 1 we can thus conclude that when taking into account the goal of non-maleficence, the best model is our final model. Which is the expected improvement as tuning the parameters per disease should give the best results.

3. Technical analysis

The most important feature of the baseline model is the striking difference between its accuracy and recall. Moderately high accuracy suggests a competent model, yet the high false negative rate means it misses most true cases of disease. This majority class collapse occurs when the model defaults to "No Findings" due to class imbalance and softmax constraints. Thus, accuracy alone is misleading, as it ignores clinically critical errors, directly relating to sub research question 1.

The accuracy drop in our continual learning model should therefore not be interpreted as a worse performance. Instead, it reflects a shift away from majority-class bias. By removing "No Finding" as an explicit output the model is forced to make binary decisions for each disease. This leads to a clinically meaningful reduction in the false negative rate. In a medical screening context, false negatives represent patients who are incorrectly considered healthy. For conditions such as pneumothorax, which requires immediate intervention, this can be deadly. The lower false negative rate is directly a reduction in patient harm, and this is the technical base for the non-maleficence argument developed in the ethics segment. While false positives increase, the potential harm is lower, they trigger additional testing rather than missed diagnoses. Since the model is meant to assist decision-making rather than act autonomously, the false positive rate is acceptable. This addresses the sub research question 3, when the two types of error have different consequences for the patients, we need to carefully decide on the metric and the threshold to reflect these consequences rather than blindly aggregate performance.

Refinement of the continual learning model includes preprocessing of the images and independently fine tuning parameters for each pathology. The per-disease fine-tuning of parameters does not generalize uniformly. When using 128×128 images, the fine-tuning led to a noticeable change in performance for only 3 of the 5 diseases. Significant improvements were observed in accuracy and false positive rate for Atelectasis, precision for Atelectasis and Infiltration, and recall and false negatives for Effusion (One sided Welch's t-test was conducted with $p < 0.05$, n = 10). These gains suggest that disease-specific parameter tuning can reduce different error types without changing the resolution.

Upscaling the images from 128 x 128 pixels to 256 x 256 pixels without any disease specific tuning also resulted in an improvement. For atelectasis, effusion and nodule the false negative rates decreased significantly. AUROC improved for effusion, infiltration and nodule, with F1 increases for effusion and nodule ($p < 0.05$). These results are consistent

with resolution being a bottleneck in the prediction of more localized pathologies. Notably, when using upscaling alone, the false positive rate did increase for some pathologies.

The final model, using both upscaling and fine tuning disease parameters achieves the strongest results overall. With recall above 0.72 across all five pathologies and false negative rates below 0.28, compared to the very low recall of the base model. Statistically significant improvements over the unprocessed, base continual learning model include F1, false negative rate and recall for effusion and pneumothorax, AUROC for nodule, and precision and false positive rate for infiltration ($p < 0.05$). The higher resolution provides more spatial information while per disease tuning controls the tradeoff between recall and false positive rate independently for each pathology.

The model functions as a decision-support tool for classifying thoracic diseases from chest X-rays, with doctors retaining final diagnostic authority. To ensure clinicians understand model outputs and mitigate risks like false negatives, aligning with non-maleficence, Grad-CAM visualizations are employed (Figure 2) (Ihongbe et al., 2024). Grad-CAM is used because a study comparing seven saliency methods shows that it is best at localizing pathologies (Saporta et al., 2022). This resolves the black-box nature of deep learning by overlaying intuitive heatmaps on original images, revealing which regions the model focuses on (Aasem and Iqbal, 2024). This enables doctors to verify alignments with clinical evidence, promotes trust, and reduces potential patient harm by facilitating informed override.

4. Ethical analysis

4.1 Ethical Dilemma and Framework

Artificial intelligence (AI) is increasingly used in medicine, impacting various areas of healthcare, including medical imaging. However, integrating AI systems into clinical practice raises ethical concerns. Automated diagnostic systems can affect clinical decision-making and incorrect predictions can lead to harmful outcomes for patients (Evans and Snead, 2024).

In automated thoracic disease classification, false negatives represent critical errors. These occur when a model fails to detect an existing disease, potentially delaying diagnosis and treatment (Evans and Snead, 2024). This leads to a central ethical dilemma: should the model prioritize accuracy or should it minimize false negatives to reduce the risk of missed diagnoses?

Ethical evaluations of medical technologies often follow the principles of biomedical ethics: autonomy, beneficence, non-maleficence, and justice. In this project, non-maleficence is the most relevant, as it emphasizes the obligation to avoid patient harm (Varkey, 2021). It stresses the importance of preventing actions that could result in unnecessary suffering or risks. This is particularly important in medical AI because diagnostic errors can directly result in patient harm. In this context, harm is defined as clinical deterioration or irreversible health complications resulting from a system not detecting a serious condition (Motloba, 2019). Our analysis focuses on the principle of non-maleficence, specifically through the lens of false negatives, as they represent the most direct threat to patient safety. For instance, failing to detect a serious condition such as a Nodule can delay urgent treatment and lead to severe health complications (Aldhafeeri, 2025).

4.2 Relevant stakeholders

The ethical concerns of AI diagnostic systems affect several stakeholders in healthcare. Patients are the most directly impacted group because missed diagnosis directly conflicts with the principle of non-maleficence. Failing to detect a serious thoracic condition such as pneumothorax may result in severe complications, including respiratory failure (Kumar et al., 2024). Conversely, while false positives (false alarms) may introduce temporary anxiety or lead to unnecessary testing, these outcomes are clinically reversible (Kumar et al., 2024).

Clinicians are responsible for interpreting model outputs and making final treatment decisions (Aldhafeeri, 2025). As our model functions as a decision-support tool, the primary risk is not just the error itself, but automation bias. A clinician may rely too heavily on algorithmic predictions and overlook the false diagnoses (Goddard et al., 2012). Because continual learning models often function as "black boxes", meaning they act as input-output mechanisms while concealing the internal logic and structure from the observer, there is an increased ethical risk, as these models change over time and thus make verification and trust in the model harder (Haterly et al., 2025).

Healthcare institutions are responsible for ensuring that the implemented systems support patient safety (Aldhafeeri, 2025). From the perspective of non-maleficence, institutions must prioritize models that reduce the risk of harmful outcomes. Although high false-positive rates may increase workload and resource demand, these impacts are more manageable than the legal and clinical consequences of undetected diseases (Aldhafeeri, 2025).

4.3 Model Design Approaches

At this stage, two design approaches have been considered. The first prioritizes overall model accuracy, a widely used metric that is easy to compute and interpret (IncludeHelp, 2023). However, accuracy combines all classification errors into a single metric and does not distinguish between error types. This limitation is particularly problematic in thoracic disease detection, where datasets are often imbalanced, meaning rare conditions occur less frequently (Lin et al., 2023; Sufian et al., 2024). In our dataset, this class imbalance applies to pneumothorax and nodule, which occur less frequently (Figure 4), along with a high 10% label noise. In such environments, models can achieve high accuracy while still performing poorly on these rare but critical cases (Sufian et al., 2024). From an ethical perspective, relying solely on accuracy conflicts with non-maleficence, as it may conceal high false-negative rates behind a high overall score, allowing serious conditions to go undetected.

An alternative design approach is to prioritize minimizing false negatives by optimizing recall (sensitivity), so that as few clinical cases as possible are missed. From the perspective of non-maleficence, this is preferable because it directly reduces the risk of harm caused by delayed or missed treatment (Kumar et al., 2024; Muyoyeta et al., 2014). Although increasing recall typically leads to more false positives, this trade-off is acceptable. False positives can result in additional tests or temporary anxiety, but they do not carry the same severe health risks as missed diagnosis (Li et al., 2025; Bernstein et al., 2023). Because different error types have different clinical consequences, prioritizing recall over accuracy becomes an ethical decision rather than purely a technical one. Furthermore, since the

model will be used as a decision-support tool, clinicians remain responsible for interpreting model predictions. This allows false positives to be reviewed and filtered in practice.

4.4 Ethical Evaluation

From the perspective of non-maleficence, the primary ethical objective of our diagnostic model is to prevent patient harm by minimizing false negatives. This directly addresses sub research question 3, which examines why minimizing false negatives is essential for upholding the principle of non-maleficence. The duty of "do not harm" is most directly violated when a disease is missed, as delayed treatment can lead to severe and potentially irreversible health consequences (Kumar et al., 2024). Therefore, prioritizing recall is essential to ensure reliable detection of potential cases. While this approach may increase the number of false positives, this tradeoff is ethically justifiable within a decision-support framework. To maintain proper value alignment, the model must never be taken at face value or replace medical judgment. Instead, by integrating Grad-CAM visualizations, clinicians can thoroughly evaluate flagged cases alongside other clinical information before making the final diagnosis. By keeping human oversight in the loop, we ensure that all potential disease cases receive appropriate attention. This supports the model's role as an assistive tool that prioritizes patient safety and supports the principle of non-maleficence.

5. Conclusions

This study examined how a CNN-based thoracic disease classifier can be optimized to balance technical performance with the ethical requirement of minimizing patient harm. By comparing the given baseline model, a Continual Learning (CL) approach, and a CL model with disease-specific parameter tuning, we have demonstrated that accuracy alone is insufficient for clinical safety. The baseline model, despite achieving high accuracy, suffered from a majority-class collapse. This resulted in an unacceptably high false negative rate exceeding 70%. Thus, confirming our first research hypothesis, where in imbalanced medical datasets, sensitivity-focused metrics are more reliable indicators of a models alignment with the principle of non-maleficence.

The implementation of a CL framework using Elastic Weight Consolidation (EWC) successfully mitigated catastrophic forgetting and allowed for more clinically meaningful detection. Our results show that while a CL approach improves performance, the integration of image upscaling (256x256) and automated, per-disease parameter tuning provides the most robust solution. Our final model achieved a recall above 0.72 across all five pathologies while maintaining an average false negative rate below 0.25. Thus meeting our clinical threshold.

While limitations such as label noise and low-resolution data remain, this research highlights that the "black-box" nature of AI can be addressed through a combination of metric selection and explainability tools such as Grad-CAM. Ultimately, the transition from a convolutional neural network to a disease-specific CL approach represents a meaningful step toward safer and more ethically responsible thoracic X-ray decision support. Future work should continue to explore threshold calibration, larger datasets, and external validation to ensure these safety gains generalize across diverse clinical environments.

Reflection on AI usage

This assignment was written by the authors in a collaborative manner. AI was used in the draft report, as stated in its AI statement, and the interaction logs were submitted.

For the final report (sprint 3), all changes from the draft were made by the authors themselves. AI was only used to review the written text for grammar, spelling and reformulate some sentences. No AI tools were used to generate new text or conclusions. We chose this approach because the draft required only minor adjustments. All final decisions regarding the content were done by the authors.

References

M. Aasem and M. J. Iqbal. Toward explainable ai in radiology: Ensemble-cam for effective thoracic disease localization in chest x-ray images using weak supervised learning. *Frontiers in big-data*, 2024. doi: 10.3389/fdata.2024.1366415.

F. M. Aldhafeeri. Governing artificial intelligence in radiology: A systematic review of ethical, legal, and regulatory frameworks. *Diagnostics*, 15(18):2300, 2025. doi: 10.3390/diagnostics15182300.

M. H. Bernstein, M. K. Atalay, E. H. Dibble, A. W. P. Maxwell, A. R. Karam, S. Agarwal, R. C. Ward, T. T. Healey, and G. L. Baird. Can incorrect artificial intelligence (ai) results impact radiologists, and if so, what can we do about it? a multi-reader pilot study of lung cancer detection with chest radiography. *European Radiology*, 33(11):8263–8269, 2023. doi: 10.1007/s00330-023-09747-1.

H. Evans and D. Snead. Understanding the errors made by artificial intelligence algorithms in histopathology in terms of patient impact. *npj Digital Medicine*, 7:89, 2024. doi: 10.1038/s41746-024-01093-w.

K. Goddard, A. Roudsari, and J. C. Wyatt. Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1):121–127, 2012. doi: 10.1136/amiajnl-2011-000089.

Y. Hadhoud, T. Mekhaznia, A. Bennour, M. Amroune, N. A. Kurdi, A. H. Aborujilah, and M. Al-Sarem. From binary to multi-class classification: A two-step hybrid cnn-vit model for chest disease classification based on x-ray images. *Diagnostics*, 14(23):2754, 2024. doi: 10.3390/diagnostics14232754.

J. Haterly, A. Søgaard, A. Ballantyne, and R. Pauwels. Federated learning, ethics, and the double black box problem in medical ai, 2025. URL `https://arxiv.org/abs/2504.20656`.

I. E. Ihongbe, S. Fouad, T. F. Mahmoud, A. Rajasekaran, and B. Bhatia. Evaluating explainable artificial intelligence (xai) techniques in chest radiology imaging through a human-centered lens. *PLOS ONE*, 2024. URL `https://doi.org/10.1371/journal.pone.0308758`.

IncludeHelp. Evaluation metrics in machine learning, 2023. URL https://www.includehelp.com/articles/evaluation-metrics-in-machine-learning.aspx.

A. Kumar, P. Patel, D. Robert, K. Shamie, K. Aneesh, B. Reddy, and A. Srivastava. Accuracy of an artificial intelligence-enabled diagnostic assistance device in recognizing normal chest radiographs: a service evaluation. *PMC Open Access*, 2024. doi: 10.1093/bjro/tzae029.

Y. Li, X. Yi, J. Fu, Y. Yang, C. Duan, and J. Wang. Reducing misdiagnosis in ai-driven medical diagnostics: A multidimensional framework for technical, ethical, and policy solutions. *Frontiers in Medicine*, 12:1594450, 2025. doi: 10.3389/fmed.2025.1594450.

M. Lin, B. Hou, S. Mishra, T. Yao, Y. Huo, Q. Yang, F. Wang, G. Shih, and Y. Peng. Enhancing thoracic disease detection using chest x-rays from pubmed central open access. *PMC Open Access*, 2023. doi: 10.1016/j.compbiomed.2023.106962.

P. D. Motloba. Non-maleficence – a disremembered moral obligation. *South African Dental Journal*, 74(1):36–40, 2019. URL https://doi.org/10.17159/2519-0105/2019/v74no1a7.

M. Muyoyeta, P. Maduskar, M. Moyo, N. Kasese, D. Milimo, R. Spooner, N. Kapata, L. Hogeweg, B. v. Ginneken, and H. Ayles. The sensitivity and specificity of using a computer aided diagnosis program for automatically scoring chest x-rays of presumptive tb patients compared with xpert mtb/rif in lusaka zambia. *PMC Open Access*, 2014. doi: 10.1371/journal.pone.0093757.

A. Saporta, X. Gui, A. Agrawal, A. Pareek, S. Q. H. Truong, C. D. T. Nguyen, V.-D. Ngo, J. Seekins, F. G. Blankenberg, A. Y. Ng, M. P. Lungren, and P. Rajpurkar. Benchmarking saliency methods for chest x-ray interpretation. *Nature Machine Intelligence*, 2022. doi: 10.1038/s42256-022-00536-x.

M. A. Sufian, W. Hamzi, T. Sharifi, S. Zaman, L. Alsadder, E. Lee, A. Hakim, and B. Hamzi. Ai-driven thoracic x-ray diagnostics: Transformative transfer learning for clinical validation in pulmonary radiography. *PMC Open Access*, 2024. doi: 10.3390/jpm14080856.

B. Varkey. Principles of clinical ethics and their application to practice. *Medical Principles and Practice*, 30(1):17–28, 2021. doi: 10.1159/000509119.

6. Appendix

Metrics	config	Atelectasis	Effusion	Infiltration	Nodule	Pneumothorax
F1	p128c	0.975	0.050	0.004	0.020	0.077
	p256nc	0.997	0.000	0.314	0.002	0.868
	p256c	0.851	0.492	0.000	0.014	0.389
Accuracy	p128c	0.814	0.922	0.004	0.000	0.462
	p256nc	0.991	0.000	0.034	0.051	0.586
	p256c	0.915	0.988	0.000	0.013	0.527
FNR	p128c	0.980	0.007	0.010	0.075	0.040
	p256nc	0.986	0.010	0.337	0.002	0.908
	p256c	0.647	0.008	0.001	0.010	0.247
FPR	p128c	0.112	0.994	0.708	0.141	0.913
	p256nc	0.316	0.649	0.393	0.914	0.132
	p256c	0.668	0.999	0.897	0.850	0.701
AUROC	p128c	0.983	0.264	0.003	0.001	0.606
	p256nc	0.999	0.000	0.003	0.001	0.606
	p256c	0.941	0.914	0.000	0.007	0.679
Precision	p128c	0.509	0.990	0.009	0.000	0.606
	p256nc	0.919	0.157	0.023	0.135	0.361
	p256c	0.862	0.999	0.002	0.056	0.442

Table 1: Performance Metrics Across Configurations

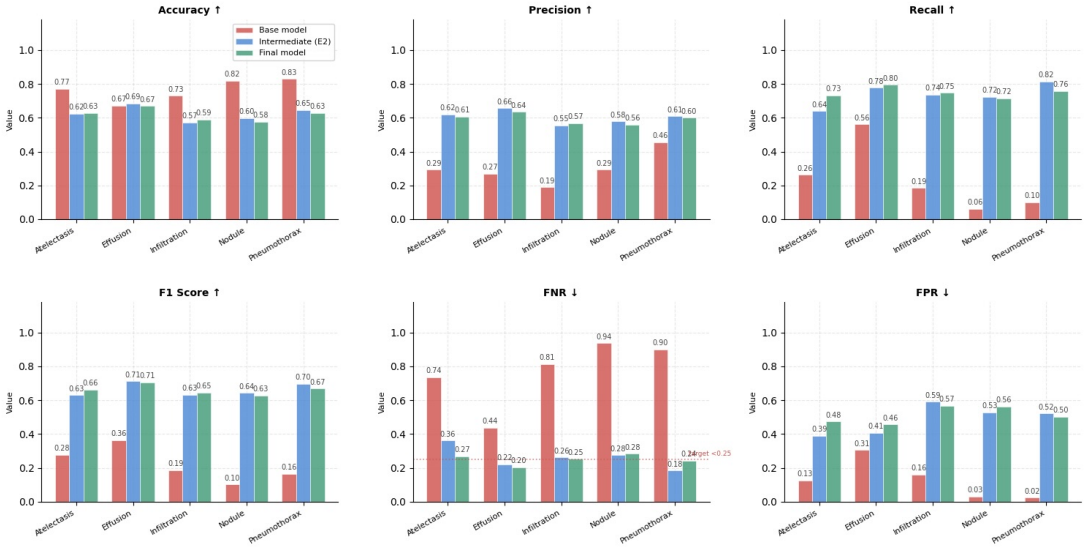

Figure 1: Barchart of Average Performance Per Model Per Disease

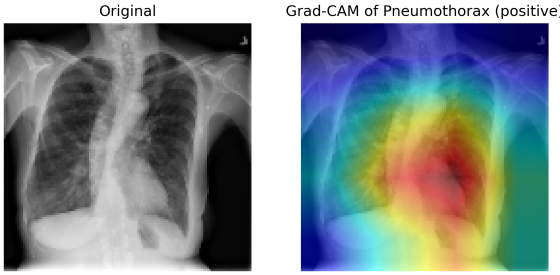

Figure 2: Chest X-ray With and Without Grad-CAM Overlay

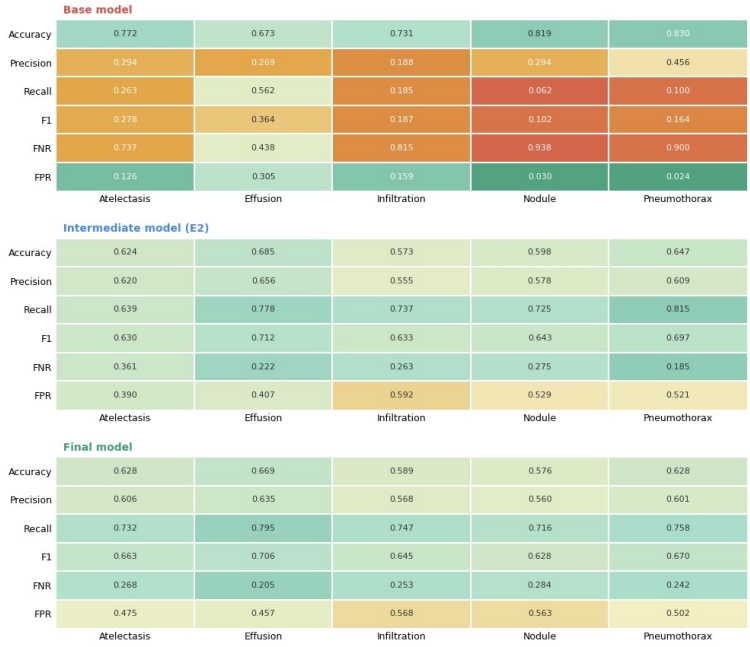

Figure 3: Heatmap of Average Performance Per Model Per Disease

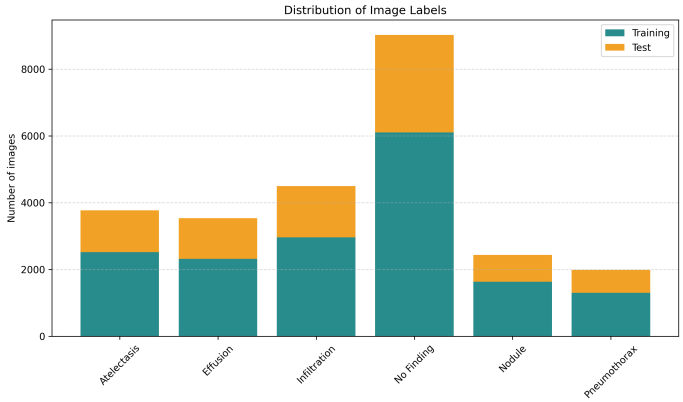

Figure 4: Bar chart of Label Distribution