CNN for detecting lung conditions on Chest X-ray for Insurance Companies

Abstract

In the insurance company sector, an issue is the presence of fraudulent claims that result in unjustified approvals. To come up with a solution, we propose the development of a Convolution Neural Network (CNN) designed to enhance the accuracy of claim validation by classifying thoracic diseases from X-rays. Our technical solution leverages a multi-layered CNN architecture for high-precision medical image analysis, serving as a reliable second opinion for insurance companies. Ethically, we prioritize transparency and manage potential conflicts of interest by ensuring that our model's decision-making processes are transparent, accessible, and aligned with the healthcare needs and fairness expected by policyholders. The template model reached an accuracy of 0.2567, whilst our first architecture reached an accuracy of 0.3184 and our second architecture 0.3145. A slight increase from the baseline model.

1. Introduction

In the insurance industry, the accuracy of diagnosing thoracic diseases is of importance for managing claims and ensuring appropriate patient treatment. In one report by an American statistician, they state that of the 800 million for annual compensation 10% would be paid as fraudulent claims. Bolton and Hand (2002) Discrepancies in diagnosis can lead to wrong payments by insurance companies and inadequate coverage for patients, often resulting in delayed access to necessary medical care. This report looks into how leveraging Convolution Neural Networks (CNN) might support and speed up the process of medical diagnosis by analyzing lung X-rays. CNN can improve diagnostic precision, providing a highly accurate second opinion that could challenge or validate initial assessments made by healthcare practitioners as mentioned by Rajpurkar et al. (2018) where they developed a CNN capable of highly accurate frontal-view chest radio-graph recognition. Such technological enhancements are not only expected to reduce the incidence of misdiagnosis but could also bring about significant cost savings. As for ethical considerations of our project we strive for inaccurate diagnoses to become less frequent, and for insurance companies to safeguard their financial resources, therefore leading to policyholders having better insurance premiums. Our main question will be how the implementation of CNN technology in medical imaging analysis will enhance the precision of thoracic disease diagnosis and therefore improve diagnostic accuracy for insurance companies.

2. Methodology

2.1 Dataset and Pre-processing Approach

The dataset comprises 112,120 frontal-view X-ray images from 30,805 unique patients, annotated with fourteen disease labels derived via natural language processing (NLP) from radiological reports. Given the NLP extraction method, label accuracy is estimated above 90%. A subset of 25,262 images is utilized, divided into 16,841 for training and 8,421 for testing, with 2,807 images reserved for external validation. The focus is narrowed to the five most frequently occurring thoracic disorders: Pneumothorax, Nodule, Infiltration, Effusion, and Atelectasis.

The pre-processing approach is divided into image normalization/standardization and data augmentation. At first, we cleaned the dataset from images that had more than 35% of fully black/white (0 or 255) pixels. Such pictures couldn't be included since, they would only distort the training process, since we have low resolution we weren't able to apply the method proposed in Robets (2021). Next, X-ray images undergo a standardization process to ensure uniformity in size and orientation K. Smelyakov1 (2022). This is accomplished by identifying the largest contour in an image to determine the object of interest, fitting a rotated bounding box around this contour, and applying rotation and cropping based on the bounding box's properties. Subsequently, images are resized to a uniform dimension (128x128 pixels), facilitating computational efficiency and consistency across the dataset.

Following standardization, the dataset is augmented to address class imbalance and increase the robustness of the model against variances in new data. Augmentation techniques include Gaussian blur, random rotation, and affine transformations (scaling, translation, and shearing). These transformations are designed to emulate realistic variations in X-ray imaging, thereby enhancing the model's generalization capabilities. The systematic review by Chlap et al. (2021) highlights the widespread adoption of data augmentation in deep learning models for medical imaging, noting its essential role in training models capable of high performance on unseen data. Furthermore, Garcea et al. (2023) emphasizes the diverse range of data augmentation techniques and their critical function in generating plausible data samples for effective model training.

The decision to over-sample minority classes (all disease classes) is justified by the aim to achieve higher accuracy on these classes. Given that these underrepresented classes correspond to critical thoracic disorders, it is crucial to ensure that the model can accurately identify them. On the other hand, the majority class (Healthy) is slightly under-sampled as it is more important for us to have better accuracy on the underrepresented classes. Figure 1 describes the difference in data classes distribution before and after data augmentation.

This part of the project did not go smoothly, we tried to implement such techniques as image overall standardization, rib suppression, image cropping to chest X-Ray, and adding contrasting filters. Since the image resolution was only 128x128, not all poorly uploaded images in the dataset made sense to be cropped. It is a whole different project to put all of the kinds of pictures we have to one standard. We tried to make a tool to brighten/darken the scan which due to too large variety of image types and lack of knowledge was way too complex to do. Application of contrast filter from a big variety such as (adaptive)-histogram equalization, hessian, CLAHE and etc from A. Giełczyk (2021) and Roberts (2021a) even degraded the accuracy of the model. Since full standardization wasn't possible, the rib suppression and image cropping were done in order to leave only lungs (put brightness to be either 0 or 255 Roberts (2021b)) tasks were eliminated after a few attempts. Only on a minority of images our algorithm gave correct output and in others it either did not find lungs, or it cropped too much and we were left with 1 or 1.5 of the lungs. Besides, we had some portion of lungs images from the side, or not lungs at all (ex: X-Ray of the gastrointestinal tract). Unfortunately, no model for lung segmentation was found, and we had no labels indicating if it was a front chest X-Ray or not, as well as we haven't had time to manually label them and create our own model to clean the images.

Augmentation was kept below 50% for every minority class. This decision was based on the principle of experimentation, allowing for adjustments in future iterations based on the performance of the model.

2.2 Research Approach

Our research leverages a CNN model optimized for classifying X-ray images into specified categories. Performance metrics including accuracy, precision, recall, F1-score, Cohen's Kappa, and the area under the precision-recall curve (AUC-PR) will evaluate the model's efficacy. The model's architecture will be refined through hyper-parameter tuning and the integration of advanced layers to improve learning capabilities and efficiency.

2.3 Template Model

Main workflow: The template model follows this workflow which is defined in `main.py`: At first, it loads and prepares the dataset. Next, the model, optimizer, and loss function are initialized. Then the model is moved to the appropriate device (GPU, CPU) for training and testing. The model generates batch data using the function in the `BatchSampler` class and iterates over the training data in each epoch. Real-time visualization of loss changes helps to monitor the training progress. After the training is done, the parameters of the model are saved in a different file in the same directory. **CNN layers:**The structure of the CNN model defined in `net.py` consists of three 2D convolution layers followed by batch normalization, ReLU activation, max-pooling, and dropout layers. Each convolution layer is defined by `nn.Conv2d`, specifying input and output channels, kernel size, and stride. Batch normalization (`nn.BatchNorm2d`) normalizes activation after each convolution. ReLU activation (`nn.ReLU`) introduces non-linearity. Max-pooling (`nn.MaxPool2d`) reduces spatial dimensions. Dropout (`torch.nn.Dropout`) prevents overfitting. **Fully Connected (Linear) Layers:**The model's linear layers consist of two fully connected layers. The first reduces the input size to 256 units with ReLU activation. The second reduces feature size to the dataset's class count for classification. **Forward Pass:** In the `forward` method, input data passes through CNN layers, flattening to a 1D tensor. This tensor passes through linear layers to produce class scores. **BatchSampler:**Functionalities include batch generation, data sampling (including balanced batches), data randomization, and handling remaining data during iteration.

2.4 Improved Models

To improve the performance of our model and better address our research question on enhancing thoracic disease diagnosis precision using CNN technology for medical imaging analysis, we have developed multiple new network architectures. These architectures have been primarily optimized according to techniques proposed by Balderas et al. (2023), focusing on optimizing network architecture. Additionally, we have implemented a novel method to systematically evaluate the performance of these network architectures. This approach allows us to identify which architecture yields the best results among the eight models. This systematic evaluation aligns with our research goal of improving diagnostic accuracy for insurance companies by enhancing thoracic disease diagnosis precision through advanced CNN technology. Furthermore, inspired by the work of Arango and Grabocka (2023), we have introduced a streamlined pipeline to optimize the code structure and reduce running time. This pipeline breaks down the entire process into smaller modular components, enhancing flexibility and facilitating easy modification or extension of specific components without affecting the overall codebase. By structuring the code in this manner, we aim to improve the efficiency and scalability of our approach.

2.4.1 Model Architecture 1 (Fig.5)

The first architecture improves the following aspects of the template model: amount of features that can be extracted by the model, and the resolution of the problem of vanishing/exploding gradients. To address the problem of vanishing gradients, residual blocks (help preserve the 'identity' function, i.e., the original input, which aids the model in learning more complex functions) are introduced into the model architecture Anwar (2019). The number of features that the model can extract is associated with the number of convolution layers, the number of filters, and their sizes. Therefore we introduce more convolution layers in the model and increase the amount of filters. To address the imbalance of the dataset even after data augmentationArvind et al. (2023), the model uses the Focal Loss Function (which has a similar idea as the Weighted Loss Function). The given model is introduced in the code as `net_ex3-4`, both files share the same idea, but in `net_ex4` the code for residual blocks is improved by rewriting the downsampling. In addition, given architecture adds more fully connected layers at the end, all these alternations in the model core are applied so the model would account for more features when making a classification. Quora (2023). In further work we conducted hyper parameters tuning, and architecture of the tuned model can be seen on Figure 9.

2.4.2 Model Architecture 2 (Fig.7)

The second architecture inherits the Focal Loss Function from the first architecture and differs from the template model by having more convolution layers with an increased number of filters and filter sizes. In the given architecture we have a gradually increasing in-depth number of filters per layer Krizhevsky et al. (2017). This is done to allow the model to extract more abstract features from the image. In addition, the architecture improves the fully connected layers in the same way as Architecture 1. Given architecture is presented in the code as `net_ex6`.

3. Experiments

In the beginning, the group started to work on data cleaning, downsampling of the overrepresented classes, and data augmentation/oversampling for underrepresented classes. In attempt to mitigate the risk of misclassification of the diagnosis. After that a new Focal Loss Function was introduced and the template model was trained. After making sure that the data augmentation and focal loss function improved the models' predictions, we started the development of different architectures by relying on existing researches.

For the model training, the group used the following approach: train the model on the first 10 epochs, then see if the loss on the test set exceeds the loss on the train set by more than 10%. In case it does, then the model starts to over-fit, and the training procedure is stopped and the model is saved. Such procedure is done in order not to mitigate the probability of misleading anyone that the model is well trained; and then give terrific results on the test set, which if deployed could potentially cost someones' life. To evaluate the model, the following metrics are computed: Accuracy (the average among all classes), Precision (helps to answer how often the positive predictions are correct, to get the picture of FP for insurance companies), Recall (explains the model's ability to expose the majority of true positive cases correctly, to get the picture of FN for patients), F-1 Score (used due to a high class imbalance as well as the high importance of FN and FP in our task; a good measure of the incorrectly classified cases), Cohen Kappa Score (assess the agreement between the predicted and true labels, accounts for imbalanced dataset thus giving objective description of the model performance).

The final step was to conduct hyper parameters tuning on the best model (Arch. 1). The hyper parameters tuning was done in two stages: firstly, tuning of learning parameters; next, tuning of convolution parameters (Atteia et al. (2022)). The parameters grid was designed with inspiration from noa (2019). The choice of parameters is based on research from those papers (ex. kernel size of filters is chosen due to its proven impact on classification outcome). For finding the best set of parameters Optuna Shekhar et al. (2022) is used.

3.1 Experimental Results

The performance of the template model is quite poor as it doesn't recognize Nodule (lb. 4) and Pneumothorax (lb. 5). Which can be due to the class imbalance. When comparing the results of architecture 1 (Fig. 4 and Tab. 1) with the template model (Fig. 2 and Tab. 1) improvement across the board is noticed. Taking into account that the data after preprocessing is still quite imbalanced; when looking at the confusion matrices we can see that Nodule and Pneumothorax are now detected way more often via Architecture 1. Which is most likely due to data augmentation and Focal loss function.

However, Cohen's Kappa score is still very low to be considered reliable even though it has improved a lot. Looking at the Fig. 4 the current issues that our model over-predicts Healthy(lb. 3) by a lot. Later we are going to apply hyper parameter tuning and optimizing to increase precision and recall and ideally achieve 0.61–0.80 Cohen's kappa score for a substantial agreement. Currently, we are looking into this Almezhghwi K (2021) article in order to improve our model and increase its performance. Accounting for both precision and recall would allow us to help both stakeholders, simultaneously.

Architecture 2 has a similar performance to Architecture 1 in terms of Accuracy, Precision, and Recall Tab. 1. There is a slight drop in F-1 and Cohen Kappa, which can be seen in confusion matrix Fig.6 where labels Atelectasis(lb. 0) and Effusion(lb. 1) are consistently over-predicted. However, it has no problem with Healthy(lb. 3) like Architecture 1 does. The aim is to analyze why exactly this happens and try to combine the best parts of these 2 architectures.

After further hyper parameters tuning of Arch. 1, we can notice improvement among all metrics 2 compare to previous models and template model. In addition we can see clearer diagonal on confusion matrix 8, which means that amount of misclassifications decreased. However, such a performance is still low, and requires better look at model architectures.

These results while improving on the template model are still unsatisfactory for the purpose of the model. Accuracy, Precision, and Recall are crucial metrics since we are dealing with life-altering outcomes. Ethics section will explore ethical threats such model can bring to both insurance companies and patients, as well as proposed resolutions.

According to the results extracted from four different models: template model, first and second architecture model and hyper tuned model, we can see a general improvement in prediction, which can be seen in Table 1 and Figures. However, model in its current state would deal harm to the patients if insurance companies would have decided to implement it. Both recall and precision scores are too low (0.34 ± 0.01) ; these 2 metrics are indicators of how good/poor job the model does in helping either insurance companies or patients. We would have to focus on one particular metric, and knowing that someones life is at stake we would consider the decrease of False Negatives (FN), of our top priority. Because right now their amount, is concerning making model unreliable. 4, 6.

4. Ethics

The deployment of Convolutional Neural Network (CNN) technology for insurance claim validation introduces complex ethical considerations, primarily due to the different interests of insurance companies and patients. Insurance companies are mainly interested in efficiency, quality, accessibility of healthcare for all and its affordability. On the other hand policyholders are interested in getting the proper care, minimum waiting lists and no unexpected costs for health care treatments. This section examines the ethical implications related to our research questions and stakeholders, identifies specific issues raised due to implemented technical features, and proposes measures to address these ethical challenges.

- **Transparency:** Due to decision-making process in CNN being a black box, produced results can be unclear, making it difficult for policyholders to understand denial reasons, thus the decision could be easily challenged Guidotti et al. (2018). This lack of transparency disproportionately impacts policyholders as it limits their ability to advocate for their interests against the insurer. Given the impact of several lung disorders which are considered in the CNN model it is of great importance that it is possible for doctors to inform patients about the use of the CNN model. Next, there is a difference between male and female X-rays, caused by the breasts of women. The radiation is slightly absorbed in the breasts, which leads to underexposure of the lungs. When a woman has had a mastectomy this could also lead to difficulties. In those X-rays a difference between both lungs occurs. This difference might be confused with lung disorders like effusion. Themes (2020) Techniques like saliency maps and activation visualizations should be used to provide insights into the CNN's focus A. Wollek (2023), while systems like LIME can can generate plain-language explanations of individual decisions, M. T. Ribeiro (2016) thus increasing stakeholder trust and facilitating informed decision-making.

- **Accountability:** Accountability processes are essential to monitor for potential misdiagnoses or unintended model biases Challen et al. (2019). We will implement the following measures to ensure accountability in our CNN model:

  * Log all model predictions and outcomes for regular auditing. Significant errors will trigger root cause analysis to identify potential issues in the training data or model architecture.
  * Establish an independent ethics review board to oversee the development, deployment, and ongoing use of the model. This board will assess potential negative impacts and recommend mitigation strategies Floridi and Cowls (2019).
  * Provide channels for stakeholders, especially policyholders, to report concerns or appeal decisions made by the model. All such reports will be thoroughly investigated and used to refine the model and associated processes Kroll (2021).

- **Conflict of Interest:** Insurance companies' main incentive is to minimize fraudulent expenses, while patients' primary value is access to high-quality healthcare. The consequences of the model producing fallacies differ. False negatives, (the model incorrectly predicts a patient as healthy) are significantly more harmful to policyholders and insurance companies as they can lead to delayed/denied care, endangering their lives Petersen et al. (2019); Bhatt et al. (2021) and given the over time evaluation of the disease this could lead to higher impact procedures (for example nodules) and causing the waiting line to grow. False positives, (the model incorrectly predicts a patient as not healthy), can result in unaccommodating expenditures for insurance companies and unnecessary stress for patients. To balance these competing interests and ensure an equitable approach, we propose the following measures:

  * Evaluate model performance with equal weighting of false positive and false negative errors to avoid bias toward insurers or policyholders, thus we made sure that our model minimizes misclassification between diseases Michael S. Klinkman (1998); Smith (2011).
  * Explicit protocols are needed to escalate borderline and contested cases to human adjudicators, with performance monitoring to audit the CNN Sukis et al. (2019). There should also be clear processes for policyholders to appeal CNN-based decisions.
  * In cases where the model results in incorrectly approving a fraudulent claim, insurance companies should have a protocol to review and potentially retrieve funds after additional verification. However, these cases must be thoroughly investigated to ensure the policyholder does not have the claimed thoracic disease before any funds are recoupedNabrawi and Alanazi (2023).

To increase understanding and trust in our CNN model's predictions, we will implement explainable AI techniques that visualize key features the model uses. This could include saliency maps highlighting regions most predictive of each disease. Additionally, a rule-based interface could show the high-level decision process, e.g. "Prediction: Pneumothorax.

Key features: right lung edge not aligned, darkening in left lung area.” Such explanations help doctors integrate the AI assistance into their diagnoses.

By prioritizing transparency, accountability, and equitability in our CNN development, we aim to create a model that responsibly serves all stakeholders.

5. Conclusions

To help out insurance companies we made a CNN model that should be able to classify thoracic diseases, providing them a second opinion on thoracic disease cases from their patients. We did this by applying CNN layers and batch normalization, ReLU activation, max-pooling, and dropout layers. After this, we also passed some hyper-parameters to try and increase the accuracy and trained our model on the dataset. As our current results are below 40% accuracy, users should not blindly trust our model to make correct decisions.

In this project, we had good ideas and their implementations for pre-processing such as poor image deletion, data standardization, and augmentation of images for imbalanced classes. However, for future work we would make better and more diverse pre-processing for the data to clean it from poorly scaled, separable images, also we would have to increase the contrast such that any formations in lungs would be better seen by the model and thus be better discriminated and classified. One of our next steps, which could improve the results, is taking a pre-trained model and retraining the model on our data by unfreezing the fully connected layers(Transfer Learning followed by Hyperparameter tuning), this option is based on the experience of Rahman (2020), which achieved great results. The argument for using of Transfer Learning is that deeper CNN models perform much better according to University of Toronto (2012), but due to computational limitations in our case, Transfer Learning could be the solution to this challenge. However, we still have to focus on the fact that our model won’t just turn into a black box, it still needs to be explainable to be transparent to the policyholders and insurance companies. Of course, we have to keep our conflict of interest regarding False Positives and Negatives, as FP are harmful to the insurance companies and FN are harmful to the policyholders,

References

Evolutionary$_{c}onvolutional_{n}eural_{n}etworks_{u}sing_{a}bc.$, 2019. Accessed: 2024-4-5.

M. T. Z. L. A. Giełczyk, A. Marciniak. Pre-processing methods in chest x-ray image classification. *PLOS ONE*, 2021. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0265949.

S. N. F. T. W. B. O. S. D. T. L. A. Wollek, R. Graf. Attention-based saliency maps improve interpretability of pneumothorax classification. *Radiology: Aritifical Intelligence*, 2023. https://arxiv.org/pdf/2303.01871.pdf.

A.-T. F. Almezhghwi K, Serte S. Convolutional neural networks for the classification of chest x-rays in the iot era. *Springer Nature*, 2021. 10.1007/s11042-021-10907-y. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8210525/.

A. Anwar. Difference between alexnet, vggnet, resnet, and inception. Toward Data Science, 2019. https://towardsdatascience.com/the-w3h-of-alexnet-vggnet-resnet-and-inception-7baaaecccc96.

S. P. Arango and J. Grabocka. Deep pipeline embeddings for automl, 2023. https://arxiv.org/pdf/2305.14009.pdf.

S. Arvind, J. V. Tembhurne, T. Diwan, and P. Sahare. Improvised light weight deep cnn based u-net for the semantic segmentation of lungs from chest x-rays. *science-direct*, 2023. 10.1016/j.rineng.2023.100929. https://www.sciencedirect.com/science/article/pii/S2590123023000567.

G. Atteia, A. A. Alhussan, and N. A. Samee. BO-ALLCNN: Bayesian-based optimized CNN for acute lymphoblastic leukemia detection in microscopic blood smear images. *Sensors (Basel)*, 22(15):5520, July 2022.

L. Balderas, M. Lastra, and J. M. Benítez. Optimizing convolutional neural network architecture, 2023. https://arxiv.org/pdf/2401.01361.pdf.

U. Bhatt, M. Andrus, A. Weller, and A. Xiang. Machine learning explainability for external stakeholders. *arXiv*, 2021. https://arxiv.org/pdf/2007.05408.pdf.

R. J. Bolton and D. J. Hand. Statistical fraud detection: A review. *Statistical Science*, 17 (3):235–249, 2002. ISSN 08834237. URL http://www.jstor.org/stable/3182781.

R. Challen, J. Denny, M. Pitt, L. Gompels, T. Edwards, and K. Tsaneva-Atanasova. Artificial intelligence, bias and clinical safety. *BMJ Quality Safety*, 28(3):231–237, 2019. https://qualitysafety.bmj.com/content/qhc/28/3/231.full.pdf.

P. Chlap, H. Min, N. Vandenberg, J. Dowling, L. Holloway, and A. Haworth. A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, 65(5):545–563, 2021. https://pubmed.ncbi.nlm.nih.gov/34145766/.

L. Floridi and J. Cowls. Establishing the rules for building trustworthy ai. *Nature Machine Intelligence*, 2019. https://doi.org/10.1038/s42256-019-0055-y.

F. Garcea, A. Serra, F. Lamberti, and L. Morra. Data augmentation for medical imaging: A systematic literature review. *Computers in Biology and Medicine*, 152:106391, 2023. https://pubmed.ncbi.nlm.nih.gov/36549032/.

R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi, and F. Giannotti. A survey of methods for explaining black box models. *arXiv:1802.01933 [cs.CY]*, 2018. https://doi.org/10.48550/arXiv.1802.01933.

O. B. E. V. K. Smelyakov1, A. Chupryna1. Standardizing cxr datasets. *Kharkiv National University of Radio Electronics*, 2022. https://ceur-ws.org/Vol-3171/paper91.pdf.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, May 2017.

J. Kroll. Outlining traceability: A principle for operationalizing accountability in computing systems. 2021. https://doi.org/10.1145/3442188.3445937.

C. G. M. T. Ribeiro, S. Singh. “why should i trust you?” explaining the predictions of any classifier. *Arxiv*, 2016. https://arxiv.org/pdf/1602.04938.pdf.

M. J. C. C. P. S. G. P. T. L. S. M. Michael S. Klinkman, MD. False positives, false negatives, and the validity of the diagnosis of major depression in primary care. *Journal of Digital Imaging*, 7:451, 1998. http://triggered.stanford.clockss.org/ServeContent?issn=1063-3987&volume=7&issue=5&spage=451.

E. Nabrawi and A. Alanazi. Fraud detection in healthcare insurance claims using machine learning. *Risks*, 11(160), 2023. https://doi.org/10.3390/risks11090160.

M. Petersen, I. A. Singh, and H. H. Meineche. Consequences of misdiagnosis: Ethical and practical challenges. *Journal of Medical Ethics*, 45(4):225–228, 2019. https://pubmed.ncbi.nlm.nih.gov/23215745/.

Quora. What are the advantages and disadvantages of fully connected layers in convolutional neural networks? https://www.quora.com/What-are-the-advantages-and-disadvantages-of-fully-connected-layers-in-convolutional-neural-network#:~:text=Higher%20Accuracy%3A%20The%20fully%20connected,predictions%20made%20by%20the%20network., 2023.

M. K. A. I. K. I. K. M. Z. K. M. K. S. Rahman, T.; Chowdhury. Transfer learning with deep convolutional neural network (cnn) for pneumonia detection using chest x-ray. *Applied Sciences*, 2020. https://doi.org/10.3390/app10093233. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8210525/.

P. Rajpurkar, J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLoS Medicine*, 15(11):e1002686, 2018. 10.1371/journal.pmed.1002686. URL https://doi.org/10.1371/journal.pmed.1002686.

D. Roberts. Applying filters to chest x-rays. *Kaggle*, 2021a. https://www.kaggle.com/code/davidbroberts/applying-filters-to-chest-x-rays.

D. Roberts. Cropping chest x-rays. *Kaggle*, 2021b. https://www.kaggle.com/code/davidbroberts/cropping-chest-x-rays.

D. Robets. Standardizing cxr datasets. *Medium*, 2021. https://www.kaggle.com/code/davidbroberts/standardizing-cxr-datasets.

S. Shekhar, A. Bansode, and A. Salim. A comparative study of Hyper-Parameter optimization tools. Jan. 2022.

N. M. L. . C. S. Smith. Causes and imaging features of false positives and false negatives on 18f-pet/ct in oncologic imaging. 2011. https://link.springer.com/article/10.1007/s13244-010-0062-3.

J. Sukis, S. Das, A. D. Wiens, F. Poursabzi-Sangdeh, L. Findlater, J. Boyd-Graber, and N. Elmqvist. Human-centered tools for coping with imperfect algorithms during medical decision-making. 2019. https://dl.acm.org/doi/10.1145/3290605.3300234.

Themes. Chest, 2020. URL https://radiologykey.com/chest-11/#:~:text=The%20major%20difference%20between%20male,not%20of%20the%20lateral%20projection.

University of Toronto. NeurIPS 2012 conference paper. 2012.

Appendix A. Computed metrics from our architectures

Metric	template model	Architecture 1	Architecture 2	Arch. 1 Hyp. Tuned
Accuracy	0.2501	0.3184	0.3145	0.3517
Precision	0.2724	0.3160	0.3143	0.3397
Recall	0.2501	0.3184	0.3145	0.3517
F1-score	0.2080	0.3129	0.2939	0.3341
Cohen Kappa	0.1081	0.1821	0.1774	0.2220

Table 1: Computed Metrics

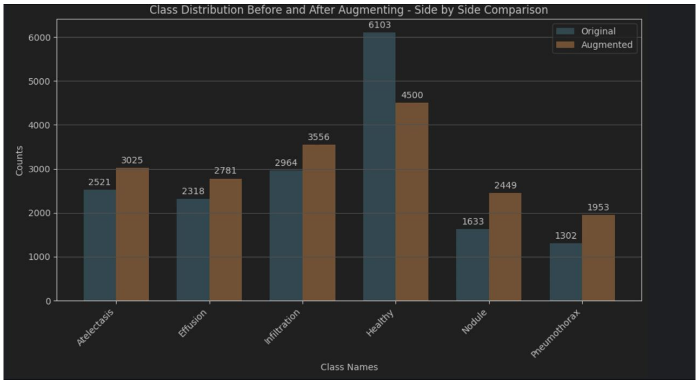

Figure 1: Augmentation Results

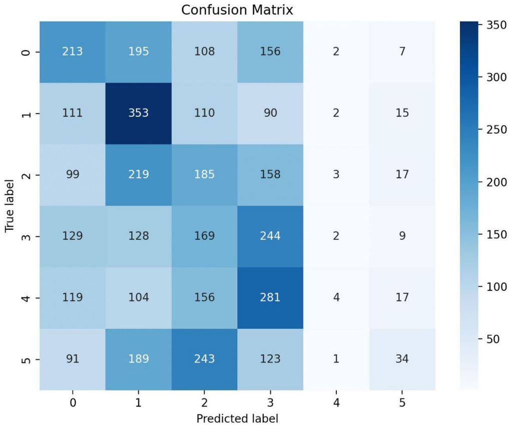

Figure 2: Confusion Matrix for Template model

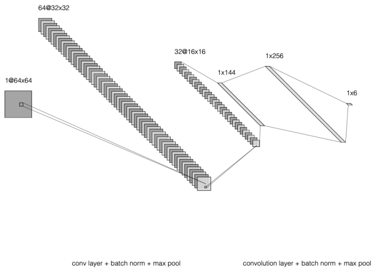
64@32x32
32@16x16
1x144
1x256
1@64x64
1x6
conv layer + batch norm + max pool
convolution layer + batch norm + max pool

Figure 3: Model Architecture template model

metric	Change Arch. 1	Change Arch. 2	Change Arch. 1 Hyp. Tuned
accuracy	+ 0.0683	+ 0.0644	+ 0.1016
precision	0.0436	+ 0.0419	+ 0.0673
recall	+ 0.0683	+ 0.0644	+ 0.1016
F1 score	+ 0.1049	+ 0.0859	+ 0.1261
Cohen's kappa score	+ 0.0740	+ 0.0693	+ 0.1139

Table 2: Change in metrics according to Table 1.

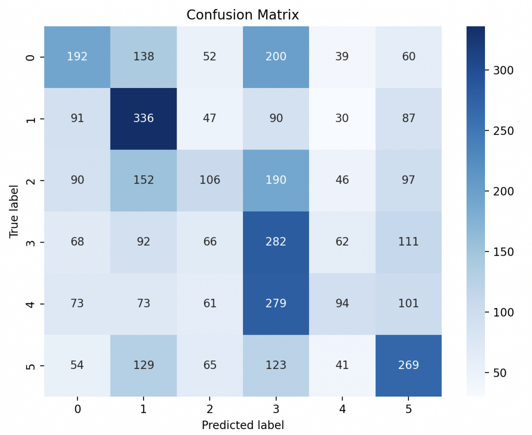

Figure 4: Confusion matrix for Architecture 1

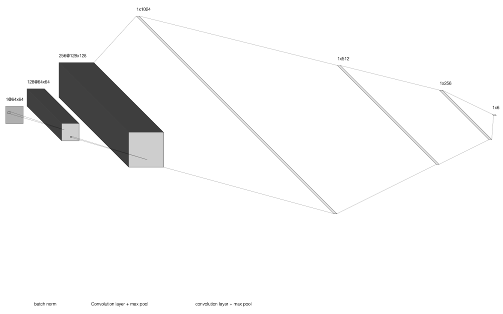

Figure 5: Model Architecture 1

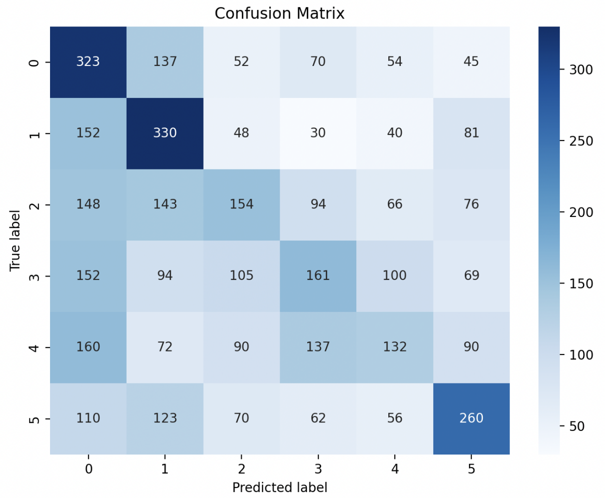

Figure 6: Confusion matrix for Architecture 2

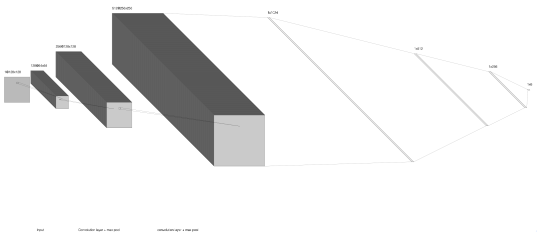

Figure 7: Model Architecture 2

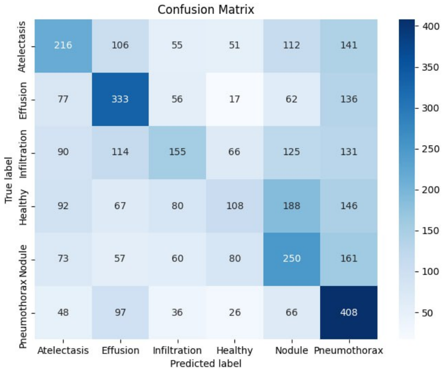

Figure 8: Confusion matrix for Architecture 1 Hyp. Tuned

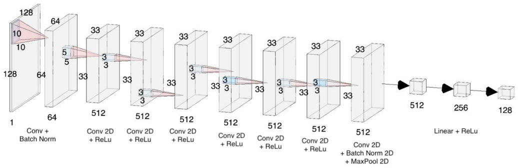

Figure 9: Model Architecture 1 Hyp. Tuned