Abnormality Detection and Predictions on Unseen Diseases in Chest X-ray Images

Minh Hoang, Varich Boonsanong, Yuxin Wu, Ludvig Liljenberg
University of Washington, Seattle, WA 98105
{minh257,varicb,xin1116,ludvil}@uw.edu

Abstract

The COVID-19 pandemic has caused enormous challenges for medical experts in identifying the abnormalities from radiography images due to the lack of pre-labeled training data. The recent need to analyze radiography images efficiently without pre-labeled unseen diseases is essential in helping combat future pandemics and reduce the burden on society. In this paper, we study the use of several Computer Vision and Deep Learning models for the tasks of abnormality detection and unseen disease prediction. Our goal is to explore the possibility of utilizing previously trained models to detect abnormal regions and predict new, unseen diseases in chest X-ray images.

1 Introduction

Experts have long warned of the potential for new diseases to emerge and cause significant health complications, particularly in regions where health resources are limited. Due to the substantial impact of such unforeseeable events on individuals and society at large, we aim to investigate the feasibility of extending the current advancement of using computer vision to predict a known disease from a given X-ray image, to also predict unknown diseases as well.

Our motivation stemmed from the spread of the novel COVID-19 virus and the bottleneck of analyzing the X-ray image. At the beginning of the pandemic, there existed no labeled training data for identifying COVID-19 from chest X-rays, so identifying COVID-19 in patients was practically impossible given the lack of data. Nevertheless, radiographers are capable of correctly separating COVID-19 X-ray images from other diseases. Thus, in this project, we try to explore the possibility of using previously trained models' intuition based on known-label diseases to extrapolate whether new images contain new diseases. We believe that if we can quantify the effectiveness of using Deep Learning models on new diseases, then we will be equipped to combat the future generation pandemic and reduce the psychological and economic burden on society.

We have seen the imbalance of hospital equipment and technology access from different parts of the country and across the continent. As a result, we come up with the second question, the robustness of these deep learning models for the spectrum of X-ray resolutions and artifacts. Specifically, after doing preliminary data exploration, we found that some hospitals printed and annotated text on the film, which can confuse the model and reduce the accuracy of their prediction. This is very important to us because, in the real world, we do not have the luxury of hand cleaned each image and preprocessing like the one that we have access to in the dataset. We simulate this by augmenting the dataset with different random noise distributions and adding some annotated characters across different locations of images.

We focus on supervised learning methods with labeled data of chest X-ray images to train multiple Deep Learning models. We perform our analyses on two types of supervised learning models. First, a binary classifier that is capable of predicting whether an image has an abnormal region or not. Second, another binary classifier that is capable of predicting whether an image has an unseen disease or not.

In summary, the main contributions of our research are:

- Conduct a case study of supervised models to determine whether or not an chest x-ray image has abnormalities and pick some of the models with the best performances.
- Use the best performing models above to investigate and quantify the effect of withholding one or more diseases from the training data on a model's performances, and investigate how well a model that has been trained on a set of diseases, performs on a previously unseen disease.
- Evaluate these model architectures and provide a guideline for selecting the best one to handle newly emerged diseases without having labeled data yet.

2 Related Works

Prior to our research, there have been several analyses on similar topics in the chest X-ray images.

Liu *et al.* [5] proposed a Segmentation-based Deep Fusion Network combining the techniques of image segmentation and deep learning for the classification of thoracic diseases. The segmentation network is trained to segment the chest X-ray images into different regions of interest (ROIs) such as the lung, heart, and clavicle. The fusion network is a deep CNN that uses the segmented ROIs as input. The proposed model achieved a mean AUC score of 0.815 that outperformed fine-tined DenseNets and all other available approaches. Moreover, the segmentation model could successfully identified 95.84% lung regions without false positive/negative segmentation. One possible limitation would be that the model was trained on a specific dataset and may not perform as well on images from different sources due to variations in image quality or other factors.

Awan *et al.* [2] applied a combination of Apache Spark framework and a Deep Transfer Learning pipeline on multiple pre-trained CNN architectures, to detect COVID-19 symptoms. They performed experiments for 2 different tasks: binary classification and 3-class classification, including normal, COVID-19 and pneumonia. For all models trained, the binary classification task achieved the accuracy of 100% and the 3-class classification task achieved more than 97% accuracy, which outperformed all other approaches. However, there are still some limitations in this proposal. Since the authors combined multiple sources of chest X-ray images as their dataset, and the overall size of the dataset is relatively small, there are some biases among the dataset, especially for binary classification.

Albahli and Yar [1] presented a multilevel classification approach using DL to diagnose COVID-19 and other chest disorders. Their approach was evaluated using different pre-trained DL models such as ResNet-50, NasNetLarge, Xception, InceptionV3, and InceptionResNetV2. The results show that ResNet-50 performed best with an average accuracy of 71.905% for COVID-19 identification and 66.634% for other diseases. Classifying X-ray images at two different stages could allow a faster speed to detect COVID-19-specific X-ray images than X-rays of other chest diseases during the pandemic. However, as the data for patients with COVID-19 is much smaller compared to other diseases, there may be some biases when training.

Cai *et al.* [4] proposed an unsupervised learning framework called Dual-distribution Discrepancy for Anomaly Detection to utilize the task of anomaly detection in both labeled and unlabeled chest X-ray images. Their framework consists of 2 modules, with Module A being trained on both labeled and unlabeled images to detect anomalous regions, and Module B models the distribution of only labeled images and is expected to show high uncertainty on unseen anomalies from Module A. The structures of the modules is an ensemble of multiple reconstruction models, and can be any variants of autoencoders (AEs). They also proposed new methods to calculate reconstruction errors and anomaly scores that consider the discrepancy between different distributions, which improve the modules' capability of discriminating anomalies. Results of the AUC values showed that the proposed framework using 3 different AEs outperformed other methods by at most 14.6% on Chest X-ray8 and 15.1% on VinDr-CXR [7] - another fairly big dataset on chest X-rays.

3 Data Collection

All the works in this project are done using 2 datasets:

Firstly, we use the NIH Chest X-ray14 dataset [9] as our training and validation data. This is a publicly available medical imaging dataset of chest radiography images that is commonly used in research related to thoracic disease classification. It contains 112,120 anonymized frontal-view X-ray images of 30,805 unique patients. Each image is a 1024x1024 PNG image, and is labeled with zero or more thoracic disease labels, for a total of 14 possible disease labels. These include: Atelectasis, Consolidation, Infiltration, Pneumothorax, Edema, Emphysema, Fibrosis, Effusion, Pneumonia, Pleural-thickening, Cardiomegaly, Nodule, Mass and Hernia. All images from the same patient only appear in either the training/validation or testing set.

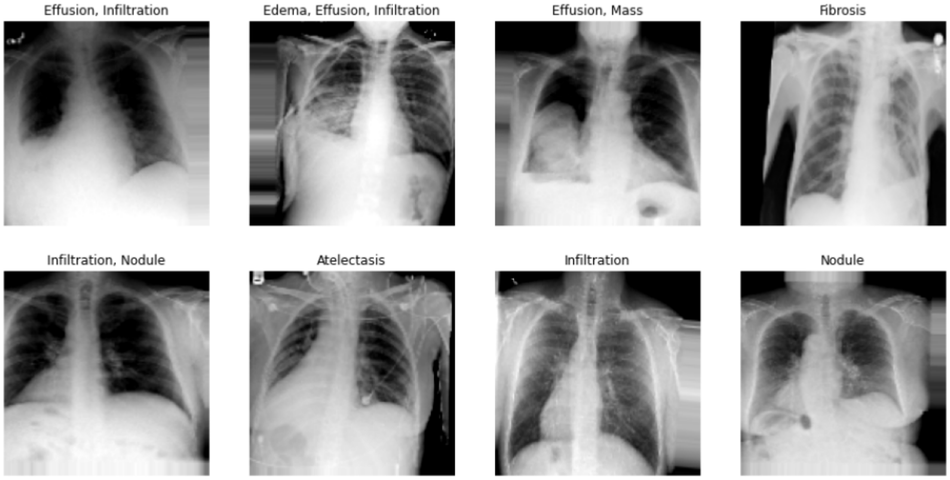


Figure 1: *Sample images of the Chest X-ray14 dataset, along with their labels.*

When evaluating our models on unseen diseases, we use another dataset called the COVIDx CXR-2 dataset [8]. This is a dataset of frontal-view X-ray images containing 29986 images from 16648 patients, Each image is also a 1024x1024 PNG image, with labels indicating whether the patient has COVID-19 (Positive) or not (Negative).

To greatly reduce the computation needed for training and validation, we resize all 1024x1024 images in both NIH ChestX-ray14 and COVIDx CXR-2 to fit each of our models' desired input size, using bi-linear interpolation. We also apply data augmentation methods, as well as normalizing the mean and standard deviation of both datasets to avoid overfitting.

4 Exploratory Data Analysis

Before conducting our analyses, some data manipulation are necessary. Since a patient may have multiple diseases in the Chest Xray14 dataset, exploratory methods such as Pricipal Component Analysis (PCA) or t-distributed Stochastic Neighbor Embedding (t-SNE) may not work well, as these methods may have trouble clustering patients with multiple diseases into one specific cluster. Table 1 shows the number of times each disease occurs in our dataset.

Label	Count
No Finding	60361
Infiltration	19894
Effusion	13317
Atelectasis	11559
Nodule	6331
Mass	5782
Pneumothorax	5302
Consolidation	4667
Pleural Thickening	3385
Cardiomegaly	2776
Emphysema	2516
Edema	2303
Fibrosis	1686
Pneumonia	1431
Hernia	227

Table 1: *Distribution of data labels in the Chest X-ray14 dataset.*

As we can see from Table 1, overall, the Chest X-ray14 dataset is quite balanced, with 60361 images not having any disease (No Finding) and 51579 images having one or more diseases. However,

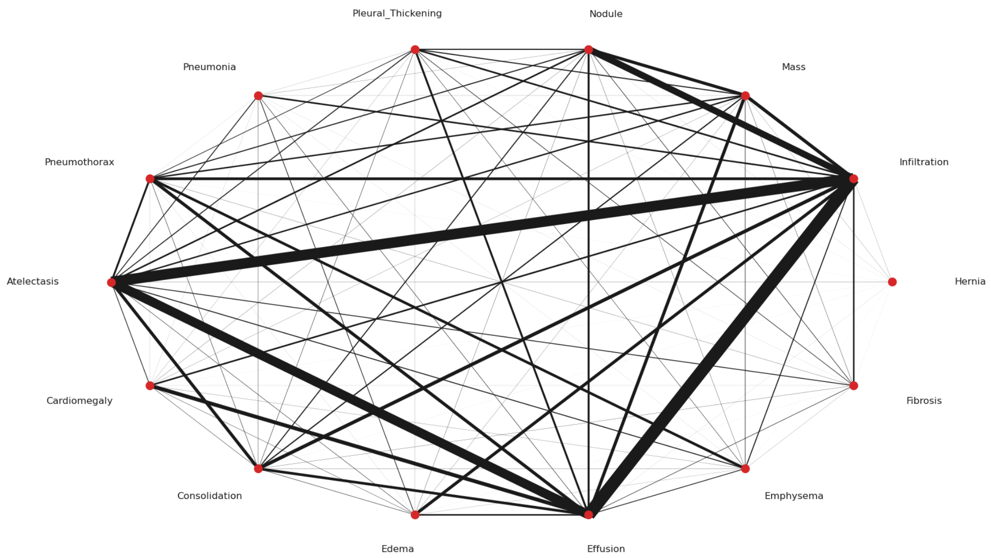


Figure 2: *Correlation between diseases in the Chest X-ray14 dataset.*

among the images with diseases, there is a huge imbalance in favor of some diseases. Specifically, Infiltration has the highest count, accounting for 38.6% and Hernia has the lowest count, accounting for only 0.4% of the data. To avoid overfitting and class imbalance among the data and improve the robustness of the deep learning model, we apply data augmentation techniques to the training data, such as random rotation, flipping, scaling and cropping.

Moreover, since there are many patients whose images show to have multiple diseases, we explore and plot the correlations of those diseases in Figure 2. According to Figure 2, there are strong correlations between Infiltration, Atelectasis and Effusion, which are the 3 diseases with the highest label counts. This means if a patient is infected with Atelectasis or Effusion, there's a high likelihood that they also suffer from Infiltration.

The COVIDx CXR-2 dataset is also quite balanced, with 15994 images of patients having COVID-19 (Positive) and 13992 images of patients not having it (Negative). For testing, the labels of the dataset are converted into binary labels using one-hot encoding.

5 Methods

We perform the following tasks and use our data to validate them:

- **Task 1:** Case Study of Abnormality Detection
- **Task 2:** Predictions on Unseen Diseases.

For **Task 1**, we want to know whether or not an image in the Chest X-ray14 dataset has regions of abnormalities. The abnormalities include the 14 labels of diseases defined in the dataset. We conduct a case study of several supervised learning models to detect abnormalities, and pick the 2 models with the best performances to perform **Task 2**. Our hypothesis is that the models with the most parameters or have the highest depths will perform better.

We use transfer learning and a *caviar strategy* to pick the models for our case study. The *caviar strategy* refers to the practice of developing and training multiple models and selecting the model having the best learning curve. Our strategy is to filter the models available from the Torchvision package, pick some of the best models with the highest Top-1 Accuracy on ImageNet, as well as having the ability to train efficiently without CUDA running out of memory midway, and conduct transfer learning on the chosen pretrained models. Following this strategy, we have picked 8 supervised learning models for our case study, all pretrained on ImageNet. These models will then be trained on the ChestX-ray14 dataset using Binary Cross Entropy loss and Adam optimizer for 10 epochs, with learning rate of 0.0002 and weight decay of 0.00002. Each model is followed by one or multiple fully connected layers, along with ReLU activation function, a dropout layer of 0.2 to prevent overfitting, and a Sigmoid layer at the end. The models are as follow:

- ResNet-50: A residual network that is 50 layers deep, formed by stacking multiple residual blocks to prevent gradient vanishing. It consists of 48 convolutional layers, 1 MaxPool and 1 AveragePool layer.
- ResNet-152: A deeper residual network with 152 layers, approximately 8 times deeper than VGG19 but still has lower complexity, with improved accuracy on both ImageNet and COCO.
- DenseNet-121: A 121-layer network where each layer is connected directly with every other layer. It consists of 4 DenseBlocks, each block has a constant dimension of the feature maps, and 3 Transition Layers to halve the number of channels.
- VGG19: The currently deepest variant of the VGG family with 19 layers, using only 3 × 3 kernels, followed by 3 fully connected layers and a Softmax layer at the end.
- MNasNet0.5: A type of neural network optimized for mobile devices, which explicitly incorporates model latency into the main objective. It mainly consists of inverted residual blocks, similar to MobileNetv2, but has less parameters.
- ResNeXt-50: A network that repeats a building block that aggregates a set of transformations with the same topology. In addition to dimensions of depth and width like ResNet-50, it also explores another essential dimension, cardinality (the size of the set of transformations).
- InceptionV3: An improved version of the Inception family by making use of the label smoothing technique, factorized 7 × 7 kernels, an auxiliary classifer to propagate label information lower down the network, and batch normalization for layers in the sidehead.
- EfficientNetV2-S: The smallest version of the EfficientNetV2 family, with better parameter efficiency and training speed compared to EfficientNet by extensively using inverted residual blocks in early layers, smaller 3 × 3 kernels and removing the last stride-1 stage.

For **Task 2**, there are 2 steps:

- First, we test our model locally using a modified Chest X-ray14 dataset: We pick one disease in the Chest X-ray14 dataset and pretend that this disease is the newly emerged disease. We remove all samples with this disease label from the dataset, and then retrain our two best performing models from Task 1 using this reduced dataset. We do this for the three diseases with the most counts (Infiltration), least counts (Hernia) and mean counts (Consolidation). To test the performance of our models, for each case, we generate a test set of 5000 images of the excluded disease, as well as images without any disease (No Finding). These images are separate from our training data. Our ideal test set would contain 2500 images of the excluded disease and 2500 images with No Finding label. However, since the number of images of is unequal among the excluded diseases, and some are less than 2500 in total, the ratio between the number of No Finding images and excluded disease images may be greater than 1:1. This is actually more realistic, as in practice, the number of people suffering a disease is supposed to be less than the number of people that do not. Each model is then evaluated on how well it is able to detect abnormalities on these unseen diseases that the models were not trained on.
- Second, we test our models to see whether they can predict a completely new disease: COVID-19. We will train our models on the whole Chest X-ray14 dataset and use the COVIDx CXR-2 dataset as our test set. Our test set contains 29986 samples of both positive or negative COVID-19 x-ray images. Each model is then evaluated on how well it is able to detect abnormalities on the COVID-19 patients.

For each task, our models are evaluated using several metrics:

- **Precision:** Precision is defined as the number of observations that are correctly classified as true positive over the total number total positive. As an example, a model with 70% precision when predicting an image to be abnormal means that it is correct 70% of the time.

$$Precision = \frac{TP}{TP + FP}$$

- **Recall:** Recall (or sensitivity) is defined as the number of observations correctly classified as true positive over the total number of true positive and false negative instances. Thus, a model with 60% recall means it correctly identifies 60% of all abnormal images.

$$Recall = \frac{TP}{TP + FN}$$

- **F1-Score:** F1-score is the harmonic mean of precision and recall metrics.

$$F1\text{-}Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

- **Matthew's Correlation Coefficient (MCC):** MCC is defined as the correlation between the predicted observations and their respective ground truths, and is a very essential metric to consider when the dataset is heavily imbalanced.

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

- **ROC-AUC**: ROC-AUC is the area under the receiver operating characteristic (ROC) curve. The ROC curve is the plot visualizing the tradeoff between true positive rate (TPR) and false positive rate (FPR).

6 Results

6.1 Task 1: Case Study of Abnormality Detection

Table 2 shows the accuracy and loss of each model after training for 10 epochs. According to Table 2 and Figure 3, VGG19 achieves the highest training accuracy, and thus the lowest training loss, but suffers overfitting heavily, as can be seen in Figure 3. This can also be explained by its low validation and testing accuracy compared to other models. The same happens with MNasNet0.5. This also explains why these 2 models' AUC are lowest (0.69, while all others are above 0.7).

Model	Accuracy			Loss	
	Training	Validation	Testing	Training	Validation
ResNet-50	0.703	0.680	0.685	0.588	0.608
ResNet-152	0.679	0.681	0.688	0.609	0.610
DenseNet-121	0.709	0.685	0.694	0.576	0.607
VGG19	**0.780**	0.640	0.639	**0.472**	0.679
MNasNet0.5	0.701	0.621	0.621	0.582	0.649
ResNeXt-50	0.714	0.690	0.692	0.567	0.605
InceptionV3	0.712	**0.698**	**0.699**	0.573	0.604
EfficientNetv2-S	0.708	**0.698**	0.692	0.582	**0.594**

Table 2: *Loss and Accuracy of 8 models after training Task 1. VGG19 achieves the best accuracy, but has the second lowest validation and testing accuracy, suggesting that it has overfitted. InceptionV3 and EfficientNetv2-S achieved the best validation and testing accuracies and losses.*

Model	Precision	Recall	F1 - Score	MCC
ResNet-50	0.665	0.652	0.658	0.366
ResNet-152	0.670	0.647	0.658	0.371
DenseNet-121	0.652	**0.736**	**0.692**	0.392
VGG19	0.594	0.704	0.644	0.288
MNasNet0.5	0.682	0.344	0.458	0.243
ResNeXt-50	0.673	0.659	0.666	0.380
InceptionV3	**0.705**	0.609	0.653	**0.394**
EfficientNetV2-S	0.670	0.669	0.668	0.381

Table 3: *Evaluation of 8 models after training Task 1. InceptionV3 achieves the best Precision and MCC scores, while DenseNet-121 has the best Recall and F1-Score.*

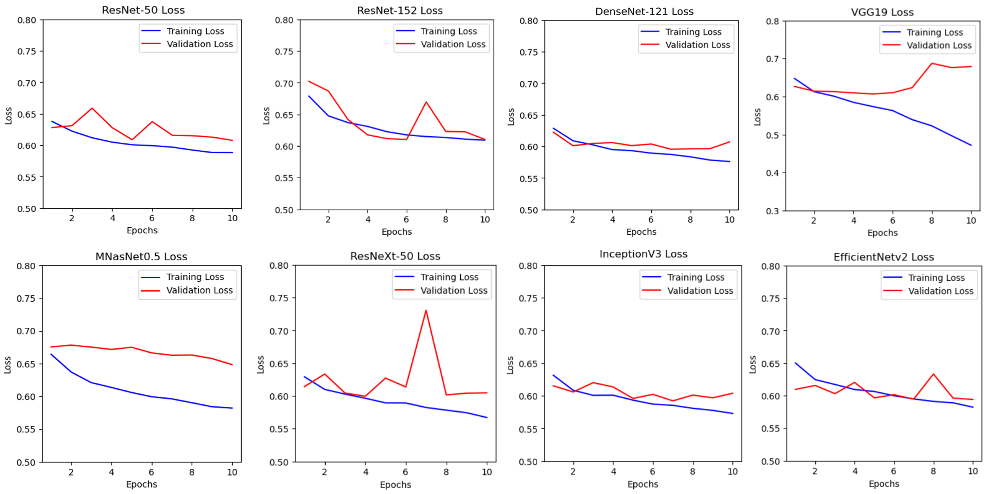


Figure 3: *Training and Validation Loss of 8 models for Task 1. VGG19 and MNasNet0.5 suffers overfitting, while ResNeXt-50, while maintaining a fairly low training loss, has an unstable validation loss.*

Using Table 2, we can identify our 2 potential model choices for Task 2: InceptionV3 and EfficientNetv2-S, as they have the highest validation and testing accuracies. However, as our training data is very big and a little imbalanced, using only accuracy and loss is not enough. We also need our evaluation metrics from Table 3. According to Table 3, InceptionV3 achieves the best Precision (0.705) and MCC scores (0.394), as well as the highest AUC value (0.75, as seen in Figure 4). For the other evaluation metrics, DenseNet-121 has the best Recall (0.736) and F1-Score (0.692). Although these 2 models have the same AUC value (0.74, according to Figure 4), EfficientNetV2-S's values in Table 3 are fairly lower, while DenseNet-121's accuracies and losses are not too worse than EfficientNetV2-S. Another thing to note is that EfficientNetV2-S has much more parameters than DenseNet-121 and InceptionV3, thus has much longer training time and higher computational costs. Therefore, based on our results, the 2 models with the best performances and will be used for Task 2 are InceptionV3 and DenseNet-121. The results also help us reject our hypothesis that models with more parameters and greater depths will perform significantly better than others.

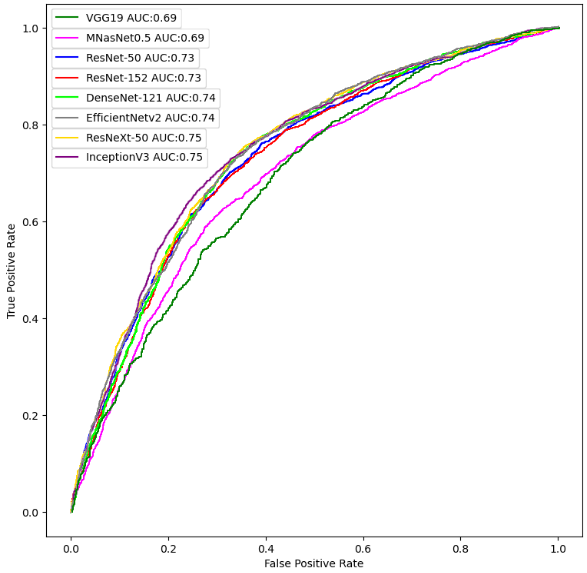


Figure 4: *ROC-AUC of 8 models aftering training Task 1. InceptionV3 and ResNeXt-50 have the best AUC, followed by DenseNet-121 and EfficientNetV2-S. VGG19 and MNasNet0.5 have the lowest AUC scores, bolstering the claim that they overfitted.*

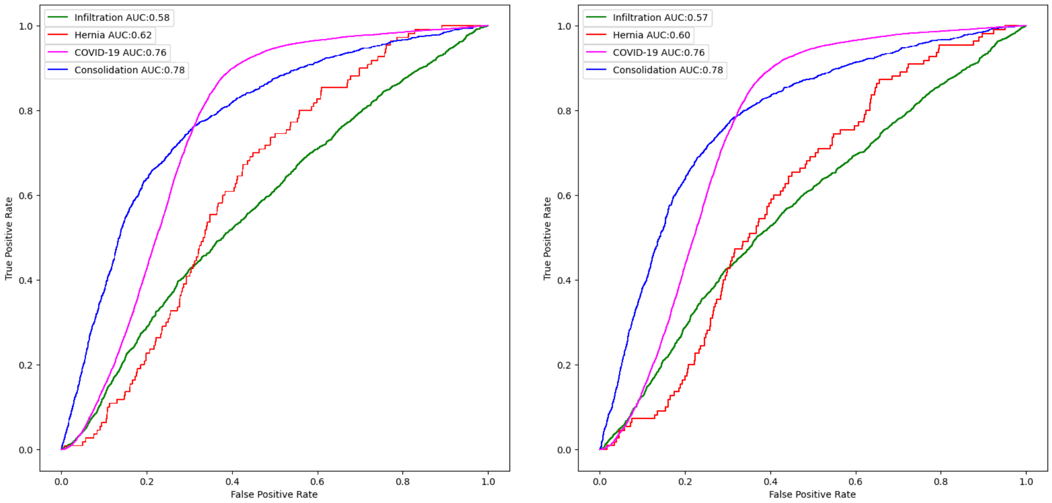
6.2 Task 2: Predictions on Unseen Diseases

Table 4 demonstrates the results of testing DenseNet-121 and InceptionV3 on predicting several unseen diseases. For our first step of **Task 2**, InceptionV3 seems to have better performances compared to DenseNet-121 when predicting unseen diseases on our modified Chest X-ray14 dataset.

Unseen Disease	Model	Accuracy	Precision	Recall	F1 - Score	MCC
Infiltration	DenseNet-121	0.528	0.582	0.196	0.293	0.073
	InceptionV3	**0.557**	**0.596**	**0.356**	**0.446**	**0.125**
Consolidation	DenseNet-121	**0.760**	**0.536**	0.634	0.581	0.417
	InceptionV3	0.734	0.495	**0.725**	**0.588**	**0.418**
Hernia	DenseNet-121	**0.710**	0.029	0.373	0.326	0.029
	InceptionV3	0.696	**0.030**	**0.409**	**0.346**	**0.036**
COVID-19	DenseNet-121	**0.755**	**0.708**	0.919	**0.800**	**0.525**
	InceptionV3	0.742	0.688	**0.943**	0.796	0.511

Table 4: *Evaluation of DenseNet-121 and InceptionV3 for Task 2. InceptionV3 has better performances on locally held out diseases (Infiltration, Consolidation, Hernia), while DenseNet-121 has slightly better results on a completely unseen disease - COVID-19.*

Both models have the worst performances in identifying Infiltration, the disease with the most counts in our dataset, achieving less than 60% accuracy and AUC, as can be seen in Figure 5. The recall values are very low, suggesting that both models have difficulties identifying samples having Infiltration (high False Negative). This leads to F1-Score and MCC scores being much less than 0.5, although the Precision and Accuracy scores are above 0.5. One possible explanation is that the abnormalities that Infiltration causes happen in a very small region of the image, and our models are not fine-tuned or strong enough to correctly identify them efficiently yet.



(a) *DenseNet-121* **(b)** *InceptionV3*

Figure 5: *ROC-AUC of DenseNet-121 and InceptionV3 for Task 2. DenseNet-121 has slightly higher AUC for Infiltration and Hernia, and the same values for Consolidation and COVID-19.*

For Consolidation and Hernia, both models achieve better evaluation scores. They can correctly classify more than 70% of the test set, and achieve fairly good AUC, especially for Consolidation (both 0.78, according to Figure 5). The Precision scores for Hernia of both models are below 0.1 due to the high imbalance between samples having Hernia and samples without diseases (ratio of roughly 1:49). While DenseNet-121's accuracy is higher than InceptionV3 for these 2 diseases, its other values are slightly lower due to the fact that it has a higher False Negative count, suggesting that it has more difficulties identifying samples having the held out disease compared to InceptionV3. Another explanation for the high testing results on Consolidation and Hernia is that the abnormality regions these 2 diseases cause is comparably larger than those caused by Infiltration.

For the second step of **Task 2**, DenseNet-121 achieves slightly better results in classifying samples with COVID-19 without training on this disease, compared to InceptionV3. Since the difference between the models' evaluation metrics are very small (less than 0.02 for all metrics), and they similar AUC (both 0.76, as seen in Figure 4), it's hard to say DenseNet-121 is a better performer than InceptionV3, as so far we have seen InceptionV3 achieving better results for most of our tasks. Nevertheless, both models have remarkably high Recall values (0.919 for DenseNet-121 and 0.943 for InceptionV3), meaning that they can identify nearly all the images having abnormalities caused by COVID-19. One possible explanation to these results is that COVID-19 causes abnormalities in multiple regions of the lung, therefore it is easier for the models to identify these regions, compared to some other diseases such as Infiltration or Mass.

7 Discussion

In **Task 1**, we conduct a case study of 8 supervised learning models on the task on detecting abnormalities caused by one or more diseases in Chest X-ray images. We hypothesize that the deeper the network is, the better it is at performing our classification task. However, through this case study, we have rejected that hypothesis. We are able to identify which models are the most suitable for our next steps (DenseNet-121 and InceptionV3), which models are not (VGG19 and MNasNet0.5, as they overfitted), and which models are potentially useful should we need more models (EfficientNetV2-S and ResNeXt-50). From the result of our study, we can see that depth and the number of parameters are not the only sole factors responsible for the performances of our models. Instead, sometimes too deep models can also lead to overfitting, which is very undesirable. Therefore, it is essential to consider multiple versions of a model family carefully so that we can avoid overfitting and underfitting our training data.

Using DenseNet-121 and InceptionV3 as our baseline models for **Task 2**, we learn that our models are able to detect some of the locally held out diseases with reasonable accuracy (Hernia and Consolidation), and struggle with some others at the same time (Infiltration) possibly due to the region caused by the diseases. Additionally, since Infiltration is highly correlated with many other disease (as can be seen in Figure 2), it makes sense that removing Infiltration from the dataset lowers the overall performance metrics of the models. Moreover, by testing out our models on a completely separate dataset of COVID-19 Chest X-ray images, we learn that the models are able to detect more than 90% of the images that are COVID-19 positive. Based on the empirical results of this task, we are confident that our 2 models DenseNet-121 and InceptionV3, with fairly similar performances, can be used to detect new and unseen diseases in Chest X-ray images in the future with limited computation resources. If we have to pick the best model for this task, InceptionV3 will be the best choice, based on its stable performances in both Task 1 and 2.

The results of **Task 2** have shown that we can now apply Deep Learning methods to detect new and unseen diseases through Chest X-ray scans of patients with high precision to detect this newly emerged disease from day 1. Moreover, given that our model works well on Chest X-ray images, it is possible that they also have the ability to predict new or unseen diseases in different areas of our body, such as brain tumor diseases through brain MRI scans or even detect abnormalities in unborn babies through abdominal MRI scans.

Finally, it should be noted that neither of our models performed perfectly. As such, while our models can be of high value to professional radiologists during diagnosis, the models should not be used as a sole tool to diagnose patients. Furthermore, our model predicts whether a Chest X-ray image contains special abnormalities. To correctly diagnose a patient to be of a specific disease, we need expert medical personnel incorporating multi-modals information (fever, breathing pattern, or medical history) that is not recorded by a radiography image.

8 Limitations and Future Works

8.1 Limitations

There are several limitations to our method and analysis that could have led us to our current results.

Even though we have good results, we believe that with access to more computing resources, we can tune our model to yield even better results. The challenges in current settings have two main aspects.

First, GPU memory is very limited when training our models. Because of this, we are unable to train very large models such as EfficientNetV2-L or MobileNetV2. Second, the size of the images and pre-processing (general artificial noises, such as Gaussian and random rotation) requires significant CPU computation and makes the process become CPU bound and not fully utilized on the GPUs.

Apart from the measurable metrics, we are assuming that our artificial noises would represent the nature of noise in real hospital settings. Given that all datasets we found have been pre-processed and that we have very limited access to the raw X-ray image, we assume that the distribution of real X-ray images would still be relatively within our simulated noises.

8.2 Future Works

There are several possible avenues we can take from our current results:

- So far, we only focus on binary classification tasks in our analysis. One possible extension is to apply our models to perform multi-label classification of multiple diseases at the same time, so that they can predict more than one unseen disease simultaneously.

- Our models successfully detect whether or not an image has abnormal regions but have yet to correctly visualize such regions. If given more time and computational resources, we can modify our models to output those abnormal regions and calculate the probability of which disease one region belongs to, should an image has multiple disease labels.

- Our model choice focuses mainly on supervised learning models, namely well-known pre-trained CNN models. In the future, we can explore different types of models, such as unsupervised learning models like Autoencoders (AEs) or Transformer-based models.

- It is possible that our models may have potential biases. Therefore, should we continue with our project, we can explore the bias in our classification models and the correlation of such bias with a specific disease. It might be possible that the current models will be biased toward a specific disease if the image happens to have a certain type of artifact that types of machines can cause.

- Our work focuses on repurposing general classification models (ResNet, EfficientNetV2), which should be able to generalize well. However, there are several X-ray-specific models, such as XNet [3], which have shown to be superior in classification tasks but may not generalize in few-shot or zero-shot learning settings [6]. We can investigate and compare our general classification models with those task-specific models in few-shot learning settings.

9 Conclusion

In this paper, we have shown with empirical results that it is possible to use transfer learning of Deep Learning models to help classify newly emerged pandemic from chest x-ray without seeing the data first. We studied multiple supervised learning models in the task of abnormality detection in radiography images. Using transfer learning and a caviar strategy, we were able to pick 8 models that are most suitable with our computation resources to perform our classification task. Of the 8 models, 2 of the best-performing ones are once again chosen to predict whether or not a Chest X-ray image contains a new or unseen disease or not. Our hypotheses and results showed that these models were able to detect an unseen disease - COVID-19 - with great precision and sensitivity. While the correlation between multiple diseases may affect the models' ability to predict one, these results are very essential to our medical fields in detecting and preventing any potential pandemic outspread in the future. We hope the success of these models will help lower the barrier of applying Deep Learning to assist diagnosis at the start of pandemic and consequently facilitate medical staff to help patient more efficiently.

Team Member Contributions:

- Minh: Conduct caviar strategy for model choice and code for model deployment for Task 1 (VGG19, MNasNet0.5, ResNeXt-50, EfficientNetV2-S), generate plots, calculate evaluation metrics, and report.
- Varich: Explanatory data analysis, examine few-shot learning, test COVID-19 as held-out disease of Task 2 and report.
- Yuxin: Code, generate plots and run tests for models (ResNet-50, ResNet-152, DenseNet-121, InceptionV3) for Task 1, and report.
- Ludvig: Data preprocessing, code for locally held-out disease training and evaluation of Task 2 and report.

References

[1] Saleh Albahli and Ghulam Nabi Ahmad Hassan Yar. Fast and accurate detection of covid-19 along with 14 other chest pathologies using a multi-level classification: Algorithm development and validation study. *Journal of medical Internet research*, 23(2):e23693, 2021.

[2] Mazhar Javed Awan, Muhammad Haseeb Bilal, Awais Yasin, Haitham Nobanee, Nabeel Sabir Khan, and Azlan Mohd Zain. Detection of COVID-19 in Chest X-ray Images: A Big Data Enabled Deep Learning Approach. *Int J Environ Res Public Health*, 18(19), September 2021.

[3] Joseph Bullock, Carolina Cuesta-Lazaro, and Arnau Quera-Bofarull. XNet: a convolutional neural network (CNN) implementation for medical X-Ray image segmentation suitable for small datasets. In *Medical Imaging 2019: Biomedical Applications in Molecular, Structural, and Functional Imaging*. SPIE, mar 2019.

[4] Yu Cai, Hao Chen, Xin Yang, Yu Zhou, and Kwang-Ting Cheng. Dual-Distribution Discrepancy for Anomaly Detection in Chest X-Rays, 2022.

[5] Han Liu, Lei Wang, Yandong Nan, Faguang Jin, Qi Wang, and Jiantao Pu. SDFN: Segmentation-based deep fusion network for thoracic disease classification in chest X-ray images. *Computerized Medical Imaging and Graphics*, 75:66–73, 2019.

[6] Feng Lu and Yunfei Liu. Image anomaly detection method based on zero-shot learning, July 14 2022. US Patent App. 17/561,869.

[7] Ha Q. Nguyen, Khanh Lam, Linh T. Le, Hieu H. Pham, Dat Q. Tran, Dung B. Nguyen, Dung D. Le, Chi M. Pham, Hang T. T. Tong, Diep H. Dinh, Cuong D. Do, Luu T. Doan, Cuong N. Nguyen, Binh T. Nguyen, Que V. Nguyen, Au D. Hoang, Hien N. Phan, Anh T. Nguyen, Phuong H. Ho, Dat T. Ngo, Nghia T. Nguyen, Nhan T. Nguyen, Minh Dao, and Van Vu. VinDr-CXR: An open dataset of chest X-rays with radiologist's annotations. *Scientific Data*, 9(1):429, Jul 2022.

[8] Maya Pavlova, Naomi Terhljan, Audrey G. Chung, Andy Zhao, Siddharth Surana, Hossein Aboutalebi, Hayden Gunraj, Ali Sabri, Amer Alaref, and Alexander Wong. COVID-Net CXR-2: An Enhanced Deep Convolutional Neural Network Design for Detection of COVID-19 Cases From Chest X-ray Images. *Frontiers in Medicine*, 9, 2022.

[9] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3462–3471, 2017.

A Appendix

Figure 6 describes the distribution of the Chest X-ray14 dataset embeddings using Principal Component Analysis (PCA). PCA fails to cluster the data points properly, due to the fact that one image may contain multiple disease labels, causing PCA to unable to choose a proper cluster for such image.

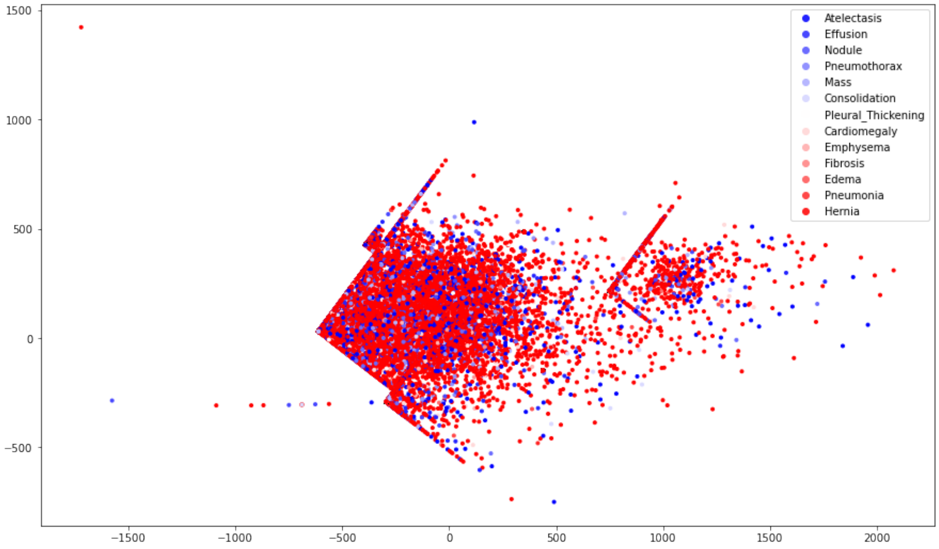


Figure 6: *Distribution of Chest X-ray14 dataset embeddings using PCA.*

Figure 7 shows sample predictions of one model from our case study in Task 1. The goal is to predict whether a Chest X-ray image is abnormal or not.

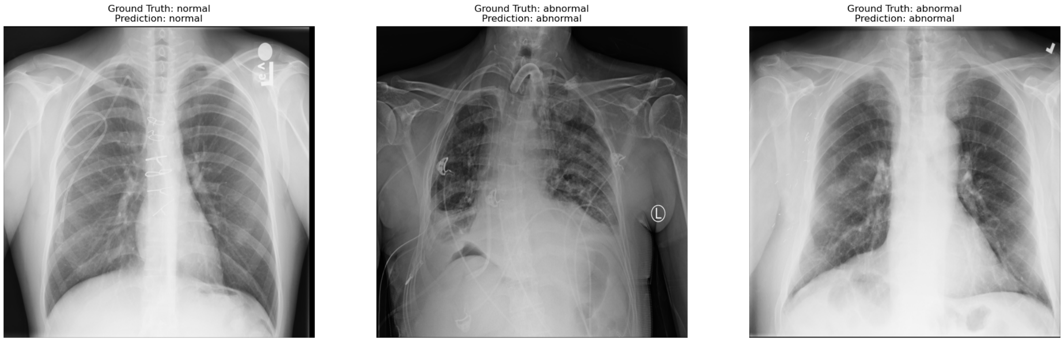


Figure 7: *Sample predictions of Task 1.*