Lung X-Ray Imaging: An Ensemble Learning Approach to CNN Disease Detection

Project Category: Image Classification of X-Rays for Lung Health
Group Alias: Project Group 11

Isidro Pride
School of Data Science
University of Virginia
Charlottesville VA, USA
zxv6mt@virginia.edu

Austin Rivera
School of Data Science
University of Virginia
Charlottesville VA, USA
atr8ec@virginia.edu

Rishi Sharma
School of Data Science
University of Virginia
Charlottesville VA, USA
bws5dk@virginia.edu

***Abstract*—Radiology has seen an escalated use of Deep Learning techniques to help identify potential disease within a patient through their medical images. Current methodologies can be expanded to be used on post-COVID-era lung imaging to not only determine if a lung is healthy or unhealthy, but further classify the suspected disease for unhealthy lungs. By having pre-labeled data from which to train, a model could recognize which features within an image are indicative of health warnings, and be paired with additional models or layers to help categorize and diagnose various diseases. There remains uncertainty around whether a single large model or an ensemble of smaller disease-specific models would prove to be the most accurate or efficient approach to developing a generalized lung disease model. The team aims to develop models that explore multiple approaches to large generalized model building, to compare their efficacy, and provide commentary on the benefits and detriments of each approach. The team also aims to develop such models for identifying specific diseases of the lungs after categorizing healthy vs. unhealthy lungs by first reviewing papers and projects that have attempted a similar approach to applying DL methods in Radiology.**

***Index Terms*—convolution neural network, radiology, lung, covid, pneumonia, tuberculosis, x-ray**

I. Literature Review

A. Current State of Deep Learning in Radiology

Deep learning has many strong applications in fields requiring image classification. Naturally this includes Radiology, where the classification of medical images can greatly expedite a diagnosis for a patient, provide healthcare providers (HCPs) with the tools to diagnose with more confidence, and detect what could be missed by the human eye. There has been an increased understanding of Neural Networks and other deep learning methods to quickly identify diseases in X-rays and MRIs. The goal of current Radiologists in this field is to spread the adoption of such models [1].

Deep learning refers to the utilization of neural networks with several hidden layers comprising the model. One heavily explored application of deep learning is computer vision, a sub-field of deep learning and machine learning that aims to interpret and classify aspects of image and video data. In building deep learning models for computer vision applications, specifically in the medical fields, there are four primary tasks with respectively more granular levels of categorization [1]:

- **Classification**: Image-level classification for identifying the presence of the class you want to detect. Returns a boolean.
- **Object Detection**: Object-level separation and classification of different known entities within an image. Returns labels and distinct boxes around each detected object within the image.
- **Semantic Segmentation**: Pixel-level classification of the objects that are the class you want to detect. Returns image with highlighted pixels for the objects of interest.
- **Instance Segmentation**: Pixel-level classification and distinction of the objects that you want to detect. Returns image with highlighted pixels of varying colors and labels for distinct objects of interest.

Within computer vision, Convolutional Neural Networks (CNNs) are the foundation of most modern deep learning models. Cheng et al. provide insight into the history and exploration into the efficacy of deeper and deeper CNNs. They discuss that while increases in depth have demonstrated improvement, they come with a cost of temporal and computational restraints that have led to the development of more efficient large models through methods such as skip connections, bottleneck blocks, multi-branch convolutions, wider networks, and ensemble networks [1]:

- **Skip Connections**: Shortcuts, or residual connections, that allow a model to skip unnecessary layers of a deep network to improve efficiency.
- **Bottleneck Blocks**: A sequence of layers that cause the model to reduce dimensionality, extract features, then restore dimensionality by reducing the number of features, increasing the kernel size, then restoring the kernel size and increasing the number of features, respectively.
- **Multi-Branch Convolutions**: Utilizing parallel convolutional layers of varying kernel sizes, then combining them via concatenation or summation, to produce a potentially

more efficient and robust model.

- **Wider Networks**: Networks with more feature maps and/or channels per convolutional layer. This is still actively being explored more as a next step to build upon these very deep networks without making them deeper. The trade off for gaining efficiency is often increased memory usage, which can be constraining.
- **Ensemble Networks**: Multiple networks within a larger network. Smaller fine-tuned networks within a larger one has the potential of being easier to train and lead to better performance.

All of these methods have shown to increase efficacy in certain applications [1], and serve to incrementally move forward the success with which computer vision has had within Radiology. However, this is not to say that there aren't challenges associated with the deep learning process in Radiology. The privacy and complexity of medical imaging can create a difficult path to creating a functioning neural network. It is important to discuss some of the primary challenges associated with this type of data and methodology.

B. An Analysis on Ensemble Learning Optimized Medical Image Classification With Deep Convolutional Neural Networks

Research has shown that there is no one size fits all model that properly and accurately classifies images of varying diseases within the medical industry. One such solution presented is to create robust ensemble models that, in the words of Muller et al. "assemble diverse models or multiple predictions and, thus, boost prediction performance" [2]. Ensemble models seek to alleviate some of the shortcomings with models in the past; namely, the fact that they have very specific use cases. By employing an ensemble mode, the medical industry hopes to create a model that generalizes to many types of images to aid medical professionals in diagnosis.

Muller et al. utilize the CHMNIST dataset comprised of healthy and cancerous colon and rectal images, COVID dataset comprised of healthy and COVID-infected lungs, ISIC dataset comprised of healthy and skin with Melanoma, and DRD dataset comprised of healthy and diabetic eyes. Some of the datasets contain x-rays whereas others contain regular high-resolution images; the point is that each disease dataset is unique. Muller et al. employ a pooled ensemble architecture comprised of DenseNet121, EfficientNetB4,VGG16, and six other pre-trained models with around 96% accuracy[2].

Despite these highly accurate results, challenges to ensemble models remain. Muller et al. leveraged the high resolution imaging of particular datasets to help train the model. Melanoma, for example, is evidenced by Asymmetry, color, and size which can be easily gleaned in a high resolution image of the skin. Not all medical datasets utilize X-rays, which don't capture color as well. Moreover, certain medical datasets are more homogeneous than the datasets used by Muller et al. Melanoma. The promising results are in ResNet101 whose network achieved 99% accuracy in classifying X-ray images[2]

C. Challenges with Methodology

There are many challenges in the current landscape of deep learning for medical images. For example, images from a single medical center can be biased toward the sampled population, which means that data from multiple centers is required to create a diverse training set [1]. This can lead to another issue regarding data security and privacy, which is a top priority for medical centers [1]. Hospitals will not freely share data of diagnostics and imaging results to be combined with other hospitals. This hampers the ability to get reliable and diverse training data for model building and testing. Good training data is essential in order to predict on unlabeled data, and being limited to a biased sample will lead to poor predictions outside of the single medical center. Cheng et al. do present a solution to the data privacy issue in the form of "Federated Learning", in which the the model itself is distributed to various medical centers to be trained, rather than said medical centers sending out their data [1]. This approach also has its hurdles, mainly with the difficulty in accounting for the diversity of populations from one medical center to the next [1].

At a higher level, there exist some issues with the CNN model when it comes to the types of images used. According to Cheng et al., the current architecture of Radiological CNNs are designed to handle 2D imaging, even if a presented image is 3D [1]. Some proposed solutions on the model include splitting 3D data into sequential 2D data, or rebuilding the CNN as a 3D architecture [1].

II. Project Proposal

A. Motivation for the Investigation

Back in 2020, the presence of debilitating, long-term lung issues increased in the US for COVID patients [3]. This means there is a need to quickly identify risk of lung disease within patient medical images. Given the progression of Deep Learning models within Radiology (see Literature Review), the team seeks to expand the use of neural networks to identify unhealthy lungs within medical images across multiple centers. Specifically, the goal is to classify healthy lungs vs. the presence of three common ailments: COVID, Tuberculosis (TB), and Pneumonia. As discussed in Section I, Radiologists have used deep learning networks, such as CNN, to identify the presence of metastases in livers [1]. The team aims to have a similar application of a CNN or other neural network model to identify the presence of aberrations in lung health.

B. Datasets

The data set is comprised of numerous sources of X-Ray lung images that are labeled with some sort of disease if unhealthy. See below the links of various X-Ray image sources in Kaggle. The sources are each healthy and unhealthy sets for each of the three lung ailments the team wishes to classify. Using this labeled training data, the team hopes to develop a reliable model to identify healthy vs. unhealthy lungs. Kaggle Links:

- Tuberculosis (TB) Chest X-ray Database (kaggle.com)
- Covid XRay Dataset (kaggle.com)
- Detecting Pneumonia in X-Ray Images — Kaggle

Each Dataset comes with different balances of healthy and unhealthy lungs:

- **Tuberculosis**: 700 w/ TB, 3500 Normal
- **COVID**: 1790 w/ COVID, 1300 Normal
- **Pneumonia**: 3875 w/ Pneumonia, 1340 Normal
- **Total Training Data**: 6365 Unhealthy, 6140 Normal

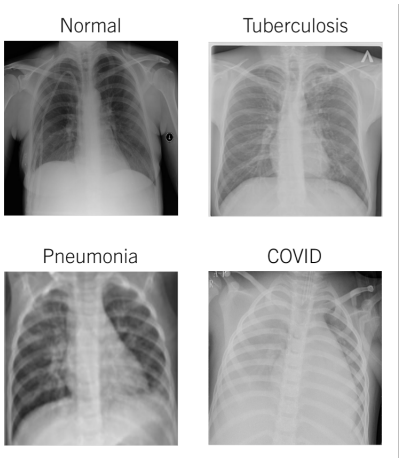

Fig. 1: Example X-Ray Images

C. Model Approaches and Experimentation

Ideally, the team would like to expand the model to identify what specific disease is present in the unhealthy lungs, measure the efficacy of different modeling approaches, and quantify the differences in temporal and computational performance metrics. Given that multiple approaches may not be feasible with the time constraints of the semester, the team is currently discussing which approach to focus on as the main model, then expanding to the others, time permitting, to compare. The three CNN-based approaches under consideration are:

1) **Ensemble Model**: The model would consist of two model layers. Layer one would be a stand alone CNN trained and validated on an even amount of healthy and unhealthy lung images, with the unhealthy being split evenly between the three diseases areas. Layer One would be tasked with identifying healthy vs. unhealthy lungs in test images. Layer two would call upon three disease-specific models, running the test image through all three of them, and tasked with identifying if the test lung shows signs of each of those diseases. This approach would allow for individualized training, tuning, and performance tracking for each of the total four sub-models within the larger ensemble. This approach leaves open the possibility for expansion into other disease ares without having to retrain the TB, COVID, and Pneumonia models. Drawbacks could be the complexity and time to work through the bugs to get these models right.
2) **Dual Model**: A one-class CNN, followed by a three-class CNN. Similar to the ensemble, this model consists of an initial model that evaluates if the test lungs are healthy or unhealthy, and would be trained in the same way. This differs in that the subsequent model would combine the three disease areas and aim to distinguish between each of them. The output of such a model could return the class with the highest probability or the probabilities of each class. This approach would simplify the process, but restricts the individual tunability of the model for the specific diseases. While these three diseases look similar to the naked eye when looking at the images, they could likely each benefit from different models. This approach also restricts expansion onto other disease areas, such as cancer, without retraining the entire model.
3) **Four-Class Model**: A single four-class CNN to classify across healthy, TB, COVID, and Pneumonia. This approach would likely result in the simplest model, but also prove to be the most restrictive when it comes to training, tuning, retraining, and interpretability.

The ensemble model is favored currently, but as mentioned above, we do see validity in evaluating the other approaches. If we were able to create all of the models, it would give us the opportunity to compare and contrast, while possibly quantifying the benefits and detriments for each of them.

D. Closing Thoughts and Considerations

Given that X-Rays can vary in quality and clarity between medical centers and data sets, the team will explore modifying the medical images using data augmentation techniques related to color adjustment, gray-scaling, and zoom. Data augmentation coupled with data sets from different sources should not only prove to help train a more robust model, but also help us address the challenges discussed above.

In the end, the hope is that the field of application for deep learning in Radiology is expanded enough where more medical centers begin adopting and endorsing the approach so as to improve the training data available for various diseases.

III. Model Architecture

The team elected to used an Ensemble model, as it has a very efficient approach to flagging unhealthy lungs, then determining what disease is present in them. This also allowed the team to delegate each disease to each team member for building preliminary models. For preliminary work, the team focused on the development of these three models so they can eventually be concatenated into a larger model that already has an existing layer to flag healthy and unhealthy lungs based on images of all 3 datasets. The concatenation of these 4 total models will be the framework of the larger final Ensemble model. The team must first finalize which individual models serve best to detect each disease.

For the diseases' preliminary architectures, each member took a different approach to develop a model, with plans to later compare these across multiple methods to guage

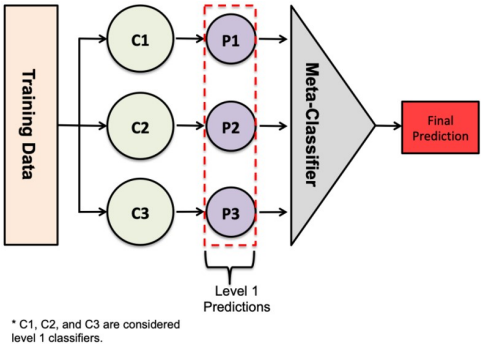

Fig. 2: Ensemble Model Architecture (High Level) [4]

performance. For TB, the preliminary model of choice was a LeNet model. A LeNet architecture is a form of CNN that has two Convolutional layers and a block of three fully-connected layers [5].

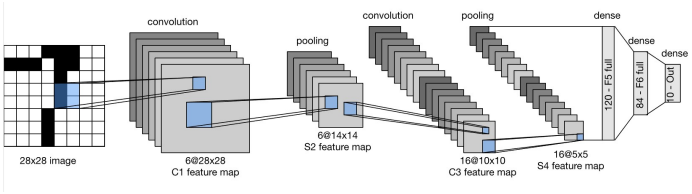

Fig. 3: A LeNet structure as described in "Dive Into Deep Learning" [5]

Describe covid approach at high level The team chose to create a model based on Residual Net 101 or ResNet101 for COVID-19 image classification. As the name suggests, ResNet101 is 101 layers deep allowing for deep feature extraction or, in other words, identifying which features are most important in classifying COVID-19-diseased lungs.

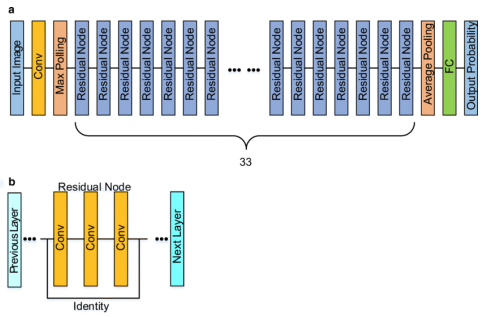

Fig. 4: ResNet101 Architecture [6]

IV. Preliminary Results

In order to obtain preliminary results, the team tested the different pieces and models that would go into creating the larger Ensemble model. Essentially, the team reviewed previous attempts at binary classification for each of the three disease data sets, and implemented the results that work best. In theory, concatenating these models within a larger model that already detects healthy vs. unhealthy lungs should results in the Ensemble model desired. The team split each of the three datasets amongst the three members, and each member then tackled creating a model for the classification of each data set.

A. Tuberculosis Model

To build a model that classifies X-Rays from the TB dataset, the team found that a LeNet network works well in both computational efficiency and accuracy. Following a similar approach to this codebook, the team developed a LeNet that could obtain a validation accuracy of 97% (see Figure 5).

```
TB_model.summary()
```

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv2d (Conv2D)              (None, 252, 252, 6)       156

max_pooling2d (MaxPooling2D) (None, 126, 126, 6)       0

conv2d_1 (Conv2D)            (None, 122, 122, 16)      2416

max_pooling2d_1 (MaxPooling2D) (None, 61, 61, 16)      0

flatten (Flatten)            (None, 59536)             0

dense (Dense)                (None, 120)               7144440

dense_1 (Dense)              (None, 84)                10164

dense_2 (Dense)              (None, 1)                 85

=================================================================
Total params: 7157261 (27.30 MB)
Trainable params: 7157261 (27.30 MB)
Non-trainable params: 0 (0.00 Byte)
_________________________________________________________________
```

Fig. 5: Preliminary TB LeNet Model Summary

To standardize the data in case more images come in from different centers, the team chose to greyscale the images at the cost of some accuracy. Greyscaling from RGB reduces the number of channels from 3 to 1, so while this improves efficiency, it comes at the cost of some accuracy (1-2%). Further adjustments to enhance the model can be explored to improve the testing accuracy within greyscaled images when developing the full Ensemble model.

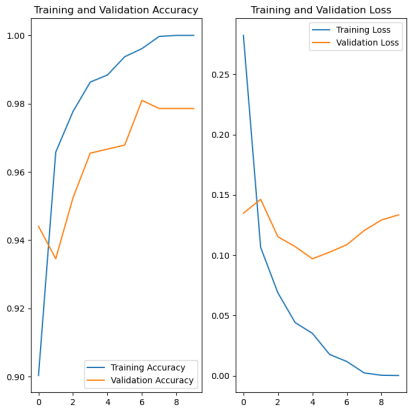

Fig. 6: Preliminary TB Model Training and Validation Loss/Accuracy

Given that our goal isn't simply to evaluate the performance of a single model, but to compare across numerous pretrained and team-developed approaches, part of our next steps will include running more computationally taxing models on the TB data to see if there is significant improvement in accuracy. Essentially, the team will evaluate if the trade-off between efficiency and performance is worth implementing in the larger Ensemble model.

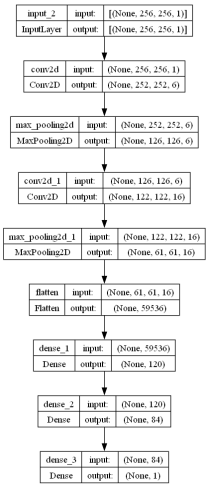

Fig. 7: Preliminary TB LeNet Model Structure

B. COVID Model

Noting the success of LeNet on Tuberculosis image classification, we experimented with the model's viability for the COVID image set as well. This yielded mixed results - the model peaked at around 70% accuracy using the same LeNet architecture that the team developed for Tuberculosis. Although these results are commendable, our goal is to achieve accuracy for COVID-19 classification that is comparable to, or exceeds, the LeNet network's results. Therefore, we explored Residual Network 101 or ResNet101 whose deep architecture allows it to excel at feature extraction. It was difficult, with our untrained eyes, to discern meaningful differences between healthy lungs and lungs affected by COVID-19 so the feature extraction abilities of ResNet101 were useful. We developed a model based on ResNet that included global averaging pooling and a sigmoid activation for the output layer as noted in our model architecture figure.

This ResNet101 trained model achieves a validation of 85% which is a notable improvement from our LeNet network evidenced in the COVID-19 Model Training and Validation Loss/Accuracy Trade-off.

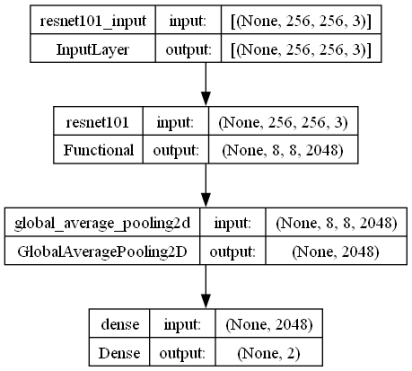

Fig. 8: COVID-19 Model Structure

Fig. 9: COVID-19 Model Training and Validation Loss/Accuracy Trade-off

C. Pneumonia Model

To develop the model for the Pneumonia dataset, before experimenting with some of the more tried and true CNN model architectures, the team wanted to get a baseline performance using a basic model architecture. In this case, the team referenced the base model from the first Codeathon for the DS 6050 course. Here we employed a model that had three convolution layers of increasing channels, max pooling layers after each Conv2D, and a single flattened then dense output layer. The intention for getting a baseline is to understand 1) if LeNet simply works that well on its own for a use case such as this one, and 2) determine if it would be worth questioning the quality of the data to evaluate if there may be unexpected identifiable artifacts in the images that give away the classification.

Layer (type)	Output Shape	Param #
input_2 (InputLayer)	[(None, 256, 256, 3)]	0
conv2d_2 (Conv2D)	(None, 256, 256, 16)	448
max_pooling2d_2 (MaxPooling2D)	(None, 128, 128, 16)	0
conv2d_3 (Conv2D)	(None, 128, 128, 32)	4640
max_pooling2d_3 (MaxPooling2D)	(None, 64, 64, 32)	0
conv2d_4 (Conv2D)	(None, 64, 64, 64)	18496
max_pooling2d_4 (MaxPooling2D)	(None, 32, 32, 64)	0
flatten_1 (Flatten)	(None, 65536)	0
dense_3 (Dense)	(None, 128)	8388736
dense_4 (Dense)	(None, 1)	129

Total params: 8412449 (32.09 MB)
Trainable params: 8412449 (32.09 MB)
Non-trainable params: 0 (0.00 Byte)

Fig. 10: Preliminary Pneumonia Basic Model Summary

With this, the basic CNN model architecture resulted in 95.7% accuracy. As a means of having a comparison to the baseline and to address our two goals stated above, the team tested the LeNet architecture on the pneumonia dataset, given its success on the TB dataset. LeNet managed a 95.5% accuracy, just slightly under-performing compared to the basic model. Though the accuracy was lower for LeNet, there were also a couple less false negative results, whereas the basic model had less false positives. While both metrics are important, we would ideally want to minimize false negatives

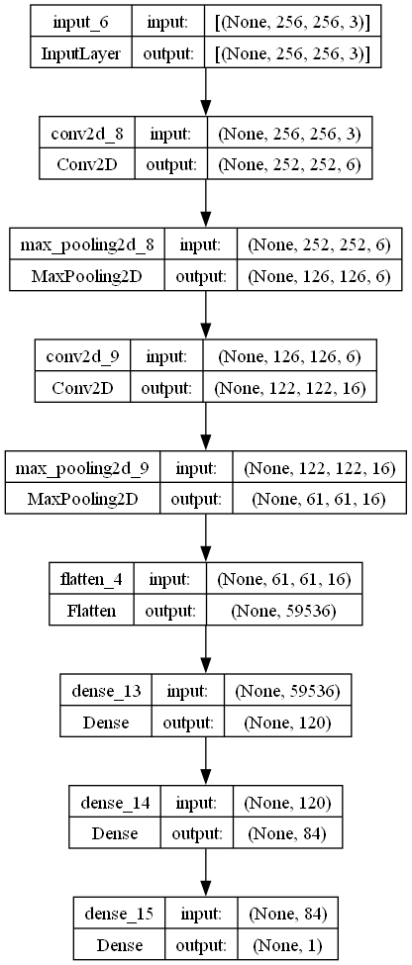

Fig. 11: Pneumonia Model Structure

over false positives. We are still exploring ways to build off of both models to determine which would be preferable or if we can create one that performs even better.

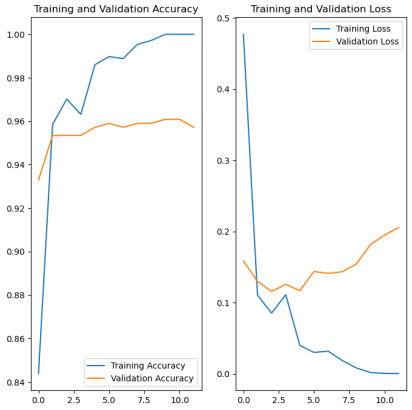

Fig. 12: Pneumonia Model Training and Validation Loss/Accuracy Trade-off

D. Summary of Preliminary Results

Having experimented on each individual model, the team put together the best results in one notebook. This will be important for eventually concatenating the models into an ensemble.

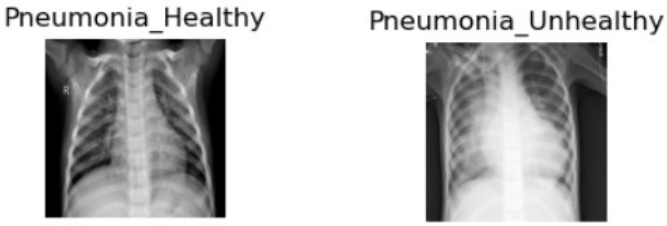

Fig. 13: Processed Healthy vs. Unhealthy Lungs Examples

Preliminary Results for Individual Models		
Disease	**Architecture**	**Accuracy**
Tuberculosis	LeNet	97.6%
COVID-19	ResNet101	85.0%
Pneumonia	LeNet	95.5%

Table 1: Best model performance for diseases.

Additional exploration into architectures that could work better after prepossessing the data will likely change the final model choice used in the larger ensemble classification.

V. Results

A. Data Preprocessing

As per the findings from preliminary research and results, the team determined that some data preprocessing can greatly benefit the models from a productions standpoint. This is due to the images requiring some degree of standardization based on how they differ in source. There are three key steps taken to complete the preprocessing for the datasets:

- **Data Size**: The number of images present in each data set differ, so the count sampled from each data set should be fixed and equal amongst each disease and healthy set.
- **Image Standardization**: Pre-processes the images by converting to grayscale, resizing, and normalizing the colors.
- **Sample Creation**: Images for each disease and healthy set are fractured into a train, validation, and test set at an 80%/10%/10% split then combined into final model-ready sets.

See in Figure 13 an example of how these the post-processed images look in their standardized form.

Some benefits of this preprocessing were immediately visible, such as the improved accuracy of the COVID model. Testing a LeNet architecture on the now processed COVID images, a validation accuracy of over 96% was achieved. It is also important to note that reduction of data set sizes seems to have negatively impacted the TB model performance, in that this particular dataset was the most diverse in terms of demographic and age groups. The reduction of such diversity likely impacted the model performance negatively. See in Table 2 the results of the updated individual models using processed datasets. These are the final models that will be used in the larger aggregate models.

Individual Models After Preprocessing Datasets		
Disease	**Architecture**	**Test Accuracy**
Tuberculosis	LeNet	96.0%
COVID-19	LeNet	96.1%
Pneumonia	Custom	95.7%

Table 2: Updated Individual Model Performances

B. Ensemble Architecture

The ensemble model architecture is similar to that in Figure 2. Essentially, there exists a first "Main" layer that determines whether an image is that of a healthy or unhealthy lung. The team chose to stick with a LeNet architecture for this binary classification as well, given the test accuracy being over 96%. From here, all images flagged as unhealthy are fed into the second layer of the ensemble. This layer consists of the three disease models obtained from post processing (see Table 2). From each of these models, a probability is returned for each disease, and the highest value will be the disease classification for that image. The second layer output accuracy is, therefore, partially dependent on the first layer accuracy. Any unhealthy images that are misclassified as healthy won't be evaluated for further specific classification.

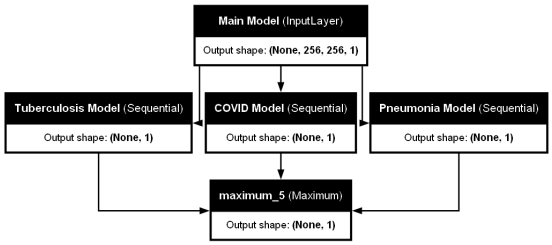

Fig. 14: Ensemble Model Architecture

C. Performance

1) Ensemble Model: The first part of the ensemble model is to identify healthy vs unhealthy lungs. As described in the Architecture, this was termed the Main model. Combining samples of each of the three disease datasets into one large sample of healthy and disease positive images, the model's sole job is to determine the probability an image is that of an unhealthy lung. Any prediction above or equal to 0.5 is flagged as unhealthy, and anything below is flagged healthy. Doing this, the model achieved a binary test accuracy of 96.4% in preliminary results.

From here, the team chose to test the complete Ensemble model for both binary accuracy (healthy vs. unhealthy) and disease classification accuracy. For better illustration, the team compared predicted vs. actual classifications for the images in a Confusion Matrix (Figure 16).

The Confusion Matrix demonstrates the ensemble model's proficiency in disease detection, however room for growth in disease differentiation. When looking at the results in more depth, it is revealed that using thresholds to decide which disease is present, while helpful, is not optimal. The most common issue was Covid or Tuberculosis being categorized as Pneumonia, due to that models slight edge in certainty. It was not uncommon for Covid model to return a value of around 0.99912, whereas the Pneumonia model would return something around 0.99991, and vice-versa, just beating out the other model. It is here where we are able to calculate the Binary Accuracy of the ensemble model, versus the Class-specific Accuracy.

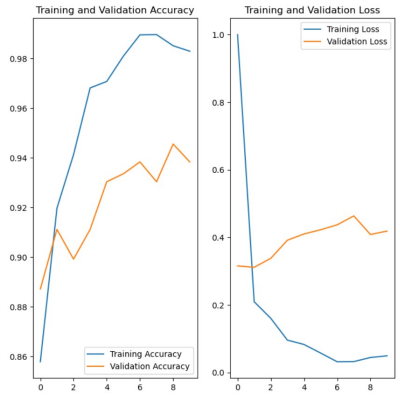

Fig. 15: Binary Healthy vs. Unhealthy Ensemble Model

To combat this issue in future work, we theorize that an additional layer of either a three-class model or three two-class models to be optionally invoked as a tie breaker, would likely prove to be effective. To the human eye, and apparently to the models, Covid and Pneumonia appear to be particularly similar, where a tie breaker model even just between them would likely result in significant reduction in misdiagnosis and an increase in Class Accuracy.

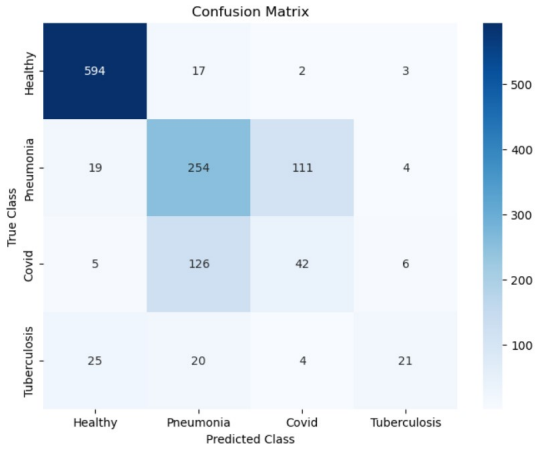

Fig. 16: Confusion Matrix Ensemble Model

2) Other Aggregate Models: In an effort to compare different methods of classification the team also tested the performance of a Dual Model and a Four Class Model. The Dual Model is simply a main model that determines healthy vs. unhealthy (similar to that in the Ensemble), and a second model that performs 3-class classification. Rather than having individual models for each disease that return probabilities, it trains on the entire aggregate unhealthy images to assign one

of three classes to each. This resulted in worse class accuracy than the Ensemble, but surprisingly had 7 less false negatives. False negatives are a major concern given that we don't want to miss anyone who indeed has a life altering condition.

The Four-Class model is a single LeNet model that performs 4-class classification. In this case, the four classes are healthy, TB, COVID, and Pneumonia. This architecture had very similar results to the Dual Model, which makes sense given that the differ only by separating the healthy images. Tables 3 and 4 depict the comparison of performance across the 3 final model types tested, each having gone through numerous iterations of layer and balance adjustments.

Outside having a few more false negatives than the other two models, the ensemble shows why having individual models adjusted for each disease can benefit in efficiency, accuracy, and flexibility. It's clear that the ensemble still has room for improvement when access to more medical centers and data becomes available. Each of the disease models, specifically Pneumonia and COVID, could benefit from adjusting to more diverse data. All of them suffer from hundreds of misdiagnoses on which disease is present, most of which occur between Pneumonia and COVID.

Final Model Accuracy Results		
Model	**Binary Accuracy**	**Class Accuracy**
Ensemble	94.3%	72.7%
Dual	92.3%	69.0%
Four Class	92.5%	68.8%

Table 3: Dual, Four-Class, and Ensemble Model Accuracy

Final Model Confusion Results					
Model	**True Positives**	**True Negatives**	**False Positives**	**False Negatives**	**Mis-diagnoses**
Ensemble	317	594	22	49	271
Dual	304	561	55	42	291
Four Class	300	562	54	40	297

Table 4: Dual, Four-Class, and Ensemble Model Confusion

VI. Conclusion

In this study, we set out to develop and evaluate an ensemble model for the classification of lung diseases using X-ray images. Leveraging deep learning architectures and data preprocessing techniques, we aimed to create a robust model capable of accurately identifying unhealthy lungs and distinguishing between different diseases, including Tuberculosis (TB), Covid, and Pneumonia.

Our preliminary research and experimentation revealed several key insights. First, data preprocessing played a crucial role in improving model performance, particularly in standardizing image sizes and colors across datasets. This preprocessing step not only significantly enhanced the accuracy and training speed of the individual models, but prepared the data to minimize underlying indicators that that would otherwise result in the combined models to differentiate based on datasets, MRIs, or hospitals. Instead the latter models that were developed were able to focus more on the feature that define the diseases.

The ensemble model architecture, consisting of a binary classification model followed by disease-specific models, demonstrated promising results in identifying unhealthy lungs and classifying diseases. The binary classification model achieved a test accuracy of 94.3% on its own, while the overall ensemble model showed decent proficiency in specific disease detection. However, there were challenges in disease differentiation, especially between COVID-19 and Pneumonia cases, leading to misdiagnosis issues.

Comparative analysis with other aggregate models, including Dual and Four-Class models, revealed that the ensemble model outperformed in terms of both binary accuracy and class accuracy. Despite its strengths, the ensemble model highlighted the need for further refinement, particularly in differentiating between similar diseases.

VII. Future Directions

There are various next steps that could be taken to not only incrementally improve the current model performance, but also to develop more optimal approaches and architectures.

A. Data and Preprocessing

- **Data Size**: Expanding the dataset and incorporating more diverse samples could enhance the model's generalization capabilities and reduce the risk of bias.
- **Data Augmentation**: Incorporating advanced data augmentation techniques and transfer learning approaches could further improve model performance, especially in scenarios with limited labeled data.
- **Data Sub-Categorization**: Diseases present themselves on a spectrum of progressive stages, and this data does not account for that level of complexity. A first step to developing this, and subsequently training different layers of models to detect early-stage versus late-stage disease, could be done by categorizing images based on how subtle or obvious they are.
- **Patient Sub-Categorization and Bias Reduction**: By looking at the images in our various datasets, it is clear from size, position, and breast and adipose tissue, that the patients images differ in age, weight, and gender and would also be worth exploring. We can tell that in the Covid and Pneumonia datasets, that most, if not all, of the patients appear to be either male or children, whereas the Tuberculosis dataset appears to contain adult women based on detection of breast tissue. It is a common issue in data science and science more broadly, that women are left out of models and studies, and so an emphasis on diversified data collection and model creation could prove to be effective methods for improving detection across patient demographics.
- **Expert Insight**: Collaboration with medical professionals and domain experts is essential to ensure the clinical relevance and validity of the model's predictions. By incorporating domain knowledge and expertise, we can develop models that not only achieve high accuracy but also contribute to improved patient outcomes and clinical

decision-making. Particularly, it would be worth getting advice on the differences between Covid and Pneumonia detection, given that the former can often induce the latter and they can exist at the same time.

B. Model and Architecture

- **Tie Breaker Models**: Additional layers or models could be introduced to serve as tiebreakers in cases where the ensemble model struggles with disease differentiation. This could involve incorporating more sophisticated algorithms or ensemble techniques to improve classification accuracy.
- **Generative Adversarial Networks**: GANs could provide a particularly elegant solution to the subtlety problem discussed about the data. Where if it is harder to detect early stage diseases, there is bound to be fewer examples to train. However, the use of GANs could allow the generation of more sample images to train on and experiment with.
- **Further Categorization**: A key to image detection is to move beyond classification, and hone in with more granularity. As we discussed in our literature review, an ideal future for the model and the initiative would be to incorporate object detection, semantic segmentation, and instance segmentation before most of the preprocessing is done. Object detection could be used to identify diseased areas of the lungs, semantic segmentation to isolate the diseased and healthy portions of the lungs, and instance segmentation to highlight the specific diseased areas. A more realistic short-term benefit that these techniques could provide would be isolation of the lungs by identifying them and masking the rest of the body to train exclusively on the lungs.

VIII. Contributions

- **Austin (Leader):** Conceived project, set success metrics, and guide model choices. Preliminary Pneumonia model development. Researched other aggregate model implementations.
- **Isidro (Data Curator):** Compiled data sets for each disease and aggregated for larger model. Data preprocessing research. Preliminary COVID model development. Researched Ensemble implementations.
- **Rishi (Organizer):** Coordinate team meetings and task completion deadlines. Match research efforts to project relevancy. Preliminary TB model development. Interpret model outputs for results. Researched current CNN landscape in Radiology.

References

[1] Cheng, P. M., Montagnon, E., Yamashita, R., Pan, I., Cadrin-Chênevert, A., Perdigón Romero, F., Chartrand, G., Kadoury, S., and Tang, A. (2021). Deep learning: An Update for Radiologists. RadioGraphics, 41(5), 1427–1445. https://doi.org/10.1148/rg.2021200210.

[2] Muller, D., Soto-Rey, I., and Kramer, F. (2022). An analysis on ensemble learning optimized medical image classification with deep convolutional Neural Networks. IEEE Access, 10, 66467–66480. https://doi.org/10.1109/access.2022.3182399

[3] Galiatsatos, P. (2022, February 28). Covid-19 Lung Damage. Johns Hopkins Medicine. https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/what-coronavirus-does-to-the-lungs

[4] Ceballos, F. (2019, September 14). Stacking classifiers for higher predictive performance. Medium. https://towardsdatascience.com/stacking-classifiers-for-higher-predictive-performance-566f963e4840

[5] Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2023, December 7). *7.6. Convolutional Neural Networks (LeNet)*. 7.6. Convolutional Neural Networks (LeNet) - Dive into Deep Learning 1.0.3 documentation. https://d2l.ai/chapter_convolutional-neural-networks/lenet.html

[6] Bangar, S. (2022, July 5). Resnet architecture explained. Medium. https://medium.com/@siddheshb008/resnet-architecture-explained-47309ea9283d