Deep Learning Based Detection of COVID-19 from Chest X-Ray Images using CNNs

1st. Hazel Wilkins
Graduate School of Arts and Sciences
Fordham University
New York, NY

2nd. Joe Margolis
Graduate School of Arts and Sciences
Fordham University
New York, NY

Abstract - This project investigates the use of deep learning models to classify chest X-ray images into three categories: Covid-19, Normal, and Viral Pneumonia. A baseline convolutional neural network (CNN) was first developed to establish reference performance, achieving moderate accuracy but showing difficulty distinguishing between pneumonia-related cases. To improve generalization and class-specific sensitivity, an EfficientNetB0 transfer learning model was implemented. To better understand how the model interprets radiographic features, multiple explainability techniques, including Grad-CAM, saliency maps, and occlusion sensitivity, were applied, confirming that the model's decisions aligned with clinically meaningful lung regions. Finally, to allow for global feature interaction, A Transformer Head was added to the EfficientNetB0 model, providing our best results with an accuracy of 98.48% and perfect recall for both sick classes. Further understanding of radiographic features were achieved by extracting token importance and implementing token heatmaps and overlays to more precisely show feature importance. These findings demonstrate the potential for transfer learning based architectures to support more reliable detection of Covid-related abnormalities in chest X-rays.

I. Introduction

The COVID-19 pandemic has placed an unprecedented strain on healthcare systems worldwide, emphasizing the need for rapid, accessible diagnostic tools. The lungs are vital organs in the human respiratory system, and any viral or bacterial infections that target them can quickly escalate into severe respiratory failure, resulting in both severe short term and long term complications. Chest X-rays are an effective method in detecting lung abnormalities and assisting physicians in assessing the severity of infections. However, relying solely on doctor-driven diagnoses introduces several challenges. The doctor-to-patient ratio is often unfavorable in low income regions, leading to delayed or missed diagnoses. Moreover, diagnoses based purely on physician experience and training can introduce bias and inconsistency across regions. The overwhelming influx of patients that present with distress in the lungs further compounds these issues, leading to clinician fatigue and decreased diagnostic accuracy. These challenges highlight the urgent need for computer-aided diagnosticians systems capable of providing timely and reliable assessments within the healthcare system.

Chest X-rays play an important but limited role in diagnosing respiratory infections because the radiographic appearances of COVID-19, viral pneumonia, and normal lungs can overlap significantly.

COVID-19 often presents with bilateral, peripheral ground-class opacities and consolidation, while viral pneumonia may show more focal or asymmetric infiltrates, and normal radiographs should exhibit clear lung fields without opacities. However, these findings are not always specific, and radiologists typically rely on a combination of symptom history, laboratory testing, and imaging to read a diagnosis. The Fleischner Society[1] [1]notes that imaging alone cannot reliably distinguish COVID-19 from other pulmonary diseases and recommends its use only in conjunction with clinical evaluation and RT-PCR testing. This diagnostic ambiguity underscores the challenge faced by automated models attempting to differentiate these conditions solely from frontal chest X-rays. Given these diagnostic limitations, machine learning and deep learning models offer a valuable opportunity to assist clinicians by identifying subtle radiographic patterns that may be too fine or inconsistent for human interpretation alone.

Manual interpretation of X-ray images is time-consuming and subject to inter-observer variability. This project aims to bridge this gap by leveraging convolutional neural networks (CNNs) to automate the classification of chest X-rays into 3 categories: COVID-19, pneumonia, or normal. This study will involve preprocessing and balancing the dataset to ensure robust model training, evaluating model performance using standard metrics to identify the most effective CNN architecture, and implementing explainability techniques to visualize model focus regions.

By developing an accurate and interpretable deep learning model, this project seeks to support health care practitioners in rapid COVID-19 screening and other respiratory infections, ultimately contributing to the growing field of AI-assisted medical diagnosis.

II. Background and Related Works

Given the still prominent impact of COVID-19 in society, there is a plethora of research into COVID-19, particularly studying alternative approaches to diagnosis. With inspiration from previous clinical use of deep learning image classification models to diagnose Pneumonia from chest X-rays, research in this process has expanded to COVID-19, showing much promise in this regard. Still, difficulty differentiating COVID-19 from other lung related ailments like Pneumonia has prevented system-wide use of image analysis for COVID-19 diagnosis, as additional research both corroborating and improving on previous findings across more datasets is still necessary.

One study, which focused solely on diagnosing COVID-19 was conducted by Akter et al., which trained 11 different types of pre-trained CNN models, settling on the MobileNetV2 model for advanced hypertuning. In the end, their model which implemented MobileNetV2 with RMSProp optimization achieved very impressive classification results, with an accuracy of 98%, a specificity of 97% and a sensitivity of 98% [2]. This project corroborated findings by Apostolopoulos et al. which also found their best results using a slightly different format of a MobileNetV2, implementing Transfer Learning and adjusting layer parameters for their CNN model. This model was also evaluated on images normalized to

[1] R. L. Rubin et al., "The Role of Chest Imaging in Patient Management During the COVID-19 Pandemic: A Multinational Consensus Statement from the Fleischner Society," Radiology, vol. 296, no. 1, pp. 172–180, Jul. 2020, doi: 10.1148/radiol.2020201365.

[2] Akter S, Shamrat FMJM, Chakraborty S, Karim A, Azam S. COVID-19 Detection Using Deep Learning Algorithm on Chest X-ray Images. Biology (Basel). 2021 Nov 13;10(11):1174. doi: 10.3390/biology10111174. PMID: 34827167; PMCID: PMC8614951.

a lower resolution than Akter et al., normalizing images to a size 200 x 266 compared to 299 x 299 [1] [3]. Even with the reduced resolution, results from this study were very similar to the prior, with an accuracy of 97.4%, specificity of 97.09%, and a sensitivity of 99.1%.

One drawback of these models is that they were trained on two class datasets, distinguishing between healthy individuals and those with confirmed COVID-19. A much more difficult task has been distinguishing between COVID-19 and other respiratory ailments, particularly Viral Pneumonia, a disease with very similar symptoms and pulmonary impact to COVID-19. For this reason, chest X-rays are rarely utilized alone in clinical practice to diagnose COVID-19, requiring other forms of COVID-19 testing to confirm any case. The current gold standard for COVID-19 diagnosis is the reverse transcription-polymerase chain reaction (RT-PCR) test, which carries its own drawbacks including being expensive, very time dependent for accurate results and time consuming to return, and difficult to attain in many communities. [4]

Lamouadene et al. aimed to fill this gap with their research, using a CNN architecture with transfer learning. Their best model used a ResNet-18 structure, achieving an accuracy of 86.2%, a sensitivity of 83.5%, and a precision of 81.3% [5]. This model performed well enough to show promise in continuing the pursuit of a model which can distinguish COVID-19 for multiclass problems, but also shows space for refinement to improve model's ability to distinguish COVID-19 from Pneumonia in patients who have an unknown classification at the time of X-ray. Our study aims to continue this research, testing other CNN architectures and implementing a Transformer component to further distinguish X-ray images of healthy, COVID-19 infected and Viral Pneumonia infected lungs.

III. Methodology

a. Dataset Acquisition and Preprocessing

The dataset used in this study was the *Covid-19 Image Dataset* hosted on Kaggle. The dataset contained 3 clinically relevant classes, Covid, Normal, and Viral Pneumonia. The images were collected from several hospital and research sources and vary in resolution, contrast, and patient positioning, making preprocessing essential. The dataset is organized into predefined train and test directories, each containing three corresponding class folders. After downloading, all dataset files were transferred to *Google Drive* to maintain consistency across training sessions in *Google Colab*.

[3] Apostolopoulos ID, Mpesiana TA. Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. Phys Eng Sci Med. 2020 Jun;43(2):635-640. doi: 10.1007/s13246-020-00865-4. Epub 2020 Apr 3. PMID: 32524445; PMCID: PMC7118364.

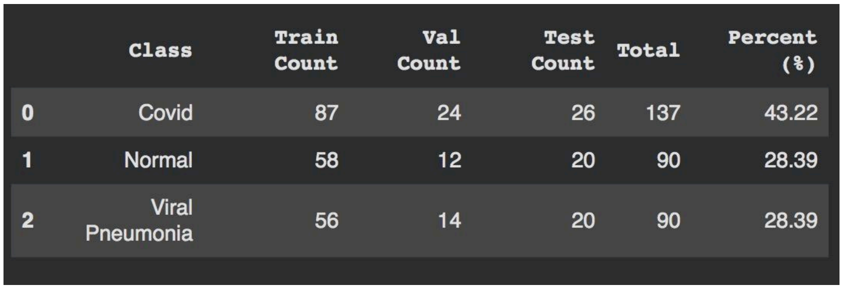
[4] Islam R, Tarique M. Chest X-Ray Images to Differentiate COVID-19 from Pneumonia with Artificial Intelligence Techniques. Int J Biomed Imaging. 2022 Dec 22;2022:5318447. doi: 10.1155/2022/5318447. PMID: 36588667; PMCID: PMC9800093.

[5] Hajar Lamouadene, Majid EL Kassaoui, Mourad El Yadari, Abdallah El Kenz, Abdelilah Benyoussef, Amine El Moutaouakil, Omar Mounkachi, Detection of COVID-19, lung opacity, and viral pneumonia via X-ray using machine learning and deep learning, Computers in Biology and Medicine, Volume 191, 2025, 110131, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2025.110131.

Data cleaning was performed to ensure high-quality inputs. Corrupted or unreadable files were removed, and all images were converted into a consistent RGB format to prevent channel-related errors during training. Directory structures were verified and standardized so that each class folder contained only valid image files. After cleaning, we confirmed that the dataset maintained balanced representation across classes within the training and test set, as visualized in Table 1.

All images were resized to 224x224 pixels to match the input requirements for both the baseline CNN and further learning models. Pixel intensities were normalized to the range [0,1] prior to training. To support model generalization and reduce overfitting, the training set was augmented using standard transformations, including random horizontal flipping, small rotations, zoom adjustments, and contrast changes. These augmentations were constrained to simulate normal variability in X-ray acquisitions, particularly imaging angles, brightness, and zoom, allowing the model to learn more robust features. Validation and test data were not augmented to ensure an unbiased performance and evaluation.

Finally, a 20% validation split was created from the cleaned training set using TensorFlow's `image_dataset_from_directory` utility with stratification based on directory structure. This produced three dataset objects: *train, validation,* and *test.* Each of these classes were shuffled and batched for efficient GPU processing. All preprocessed datasets were cached and prepared for use in the baseline, transfer learning, transformer head, and fine-tuned models that follow.



	Class	Train Count	Val Count	Test Count	Total	Percent (%)
0	Covid	87	24	26	137	43.22
1	Normal	58	12	20	90	28.39
2	Viral Pneumonia	56	14	20	90	28.39

Table 1: This table summarizes the number of images per class in each dataset split and shows the overall class balance for the Covid, Normal, and Viral Pneumonia categories. Reporting class distribution is important for understanding potential sources of bias during model training and evaluation.

b. Baseline CNN Architecture

To establish a performance benchmark for comparison with transfer learning approaches, a custom Convolutional Neural Network (CNN) was implemented and trained from scratch on the preprocessed charts X-ray dataset. The baseline architecture was designed to be lightweight, yet expressive enough to learn discriminant radiographic features directly from the data.

The model consists of three sequential convolutional blocks. Each block includes a 2D convolution layer followed by a max-pooling operation to progressively reduce spatial dimensionality while preserving salient features. The first block applies 32 filters, the second uses 64 filters, and the third uses 128 filters, all with 3x3 kernels, ReLU activations, and 'same' padding. Max-pooling layers with a stride of 2x2 downsample the feature maps after each convolutional layer, enabling hierarchy extraction of lung structure patterns, such as edges, opacities, and texture differences associated with Covid and viral pneumonia.

Following the convolutional blocks, the feature maps are flattened and passed through as a 256-unit fully connected layer with ReLU activation to learn high-level representations. A dropout rate of 0.5 is applied to mitigate overfitting, particularly given the limited data size. The final classification layer is a three-unit softmax output, corresponding to the three diagnostic categories: Covid, Normal, and Viral Pneumonia.

The model was trained using the Adam optimizer, with a learning rate of 1e-4, and sparse categorical cross-entropy as the loss function. Accuracy was tracked as the primary training metric. Early stopping and model checkpointing were utilized to prevent overfitting and ensure the best model weights were retained. The baseline architecture provides a controlled setup, enabling direct comparison with the performance gains achieved through further advanced models.

c. Transfer Learning with EfficientNetB0

To improve performance beyond the baseline CNN, we adopted a transfer learning approach using *EfficientNetB0,* a pretrained convolutional architecture that has demonstrated strong performance across a wide range of vision tasks. Rather than training the full network from scratch, the pretrained backbone was used as a fixed feature extractor while a lightweight classification head was trained on top. This allows the model to reuse rich visual representations learned from large scale natural image datasets and adapt them to the radiographic domain with relatively few parameters updated.

The overall transfer learning workflow is illustrated in Fig 1, which summarizes the frozen backbone, preprocessing pipeline, and the added fully connected layers. During the initial stage, all EfficientNet convolutional layers were frozen, and only the appended dense layers were trained. This stabilizes training and prevents early stage overfitting. After convergence, the upper portion of the network was selectively unfrozen for fine-tuning, enabling the higher-level feature maps to adapt more closely to Covid-specific radiographic patterns while keeping lower-level filters intact.

d. Fine Tuning the Pretrained Model

Following the initial transfer learning stage, a targeted fine-tuning phase was conducted to further adapt EfficientNetB0 to the chest X-ray domain. During this phase, only the upper convolutional blocks of the pretrained backbone were unfrozen, while the earlier layers remained fixed. This selective unfreezing strategy allows the model to refine highlevel semantic features, such as opacities or lung-field texture irregularities, without disrupting the low-level filters responsible for edge and gradient extraction. As illustrated in Fig 1, fine-tuning was performed with a reduced learning rate to ensure stable, incremental updates to the pretrained weights.

The fine-tuning stage uses the same data preprocessing and augmentation pipeline as the frozen-backbone training, maintaining consistency across the two phases. By updating only a subset of the model parameters, the network can specialize more effectively to the visual characteristics present in Covid positive and other X-ray categories while avoiding the overfitting risks associated with full-model retraining on a relatively small dataset. This approach is standard in medical imaging transfer learning workflows and provides an essential intermediate step between generic pretrained representations and fully task-specific feature extraction.

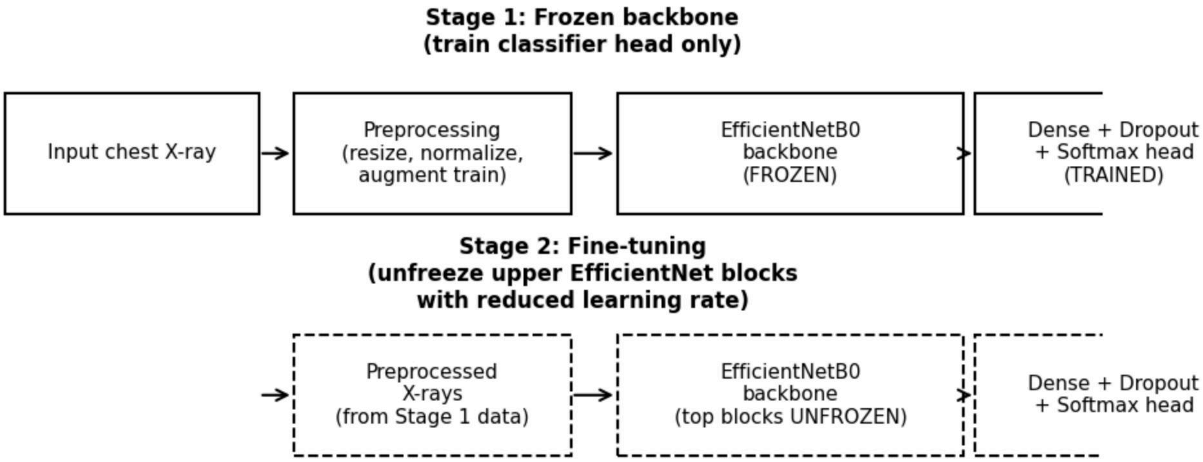

Stage 1: Frozen backbone
(train classifier head only)
Input chest X-ray
Preprocessing (resize, normalize, augment train)
EfficientNetB0 backbone (FROZEN)
Dense + Dropout + Softmax head (TRAINED)
Stage 2: Fine-tuning
(unfreeze upper EfficientNet blocks
with reduced learning rate)
Preprocessed X-rays (from Stage 1 data)
EfficientNetB0 backbone (top blocks UNFROZEN)
Dense + Dropout + Softmax head

Fig 1. Transfer learning and fine-tuning workflow. In Stage 1, input chest X-rays are preprocessed and passed through an EfficientNetB0 backbone with all convolutional layers frozen, while only the appended dense classifier head is trained. After convergence, Stage 2 selectively unfreezes the upper EfficientNet blocks and continues training with a reduced learning rate, allowing high-level feature maps to adapt to Covid-specific radiographic patterns while preserving lower-level representations.

e. *Transformer Head*

While very successful at extracting local features, one of the main constraints of Transfer Learning and CNN models are their performance with global variables, or in this context interactions between areas of the lung. To address this issue, the final model adds a Transformer Head to the EfficientNetB0 pipeline which tokenizes the local features extracted by EfficientNetB0 and applies self attention to find global interactions. Similarly to the Transfer Learning training pipeline, this model was first trained with the backbone frozen, only tuning the Transformer Head to retain the information captured in the pretrained models. This model also used the same augmented training data as the prior models along with the same validation and test sets. Due to noticed stability issues in the model's convergence, a linear warmup learning rate scheduler was applied with added epochs using TensorFlow. This process starts with a very low learning rate when the model has randomly initialized weights, then slowly increases the learning rate over each epoch until it reaches the base learning rate ($1e^{-4}$), preventing the model from converging too aggressively as it is still learning weights in the early epochs. The Transformer architecture is shown in Fig. 2.

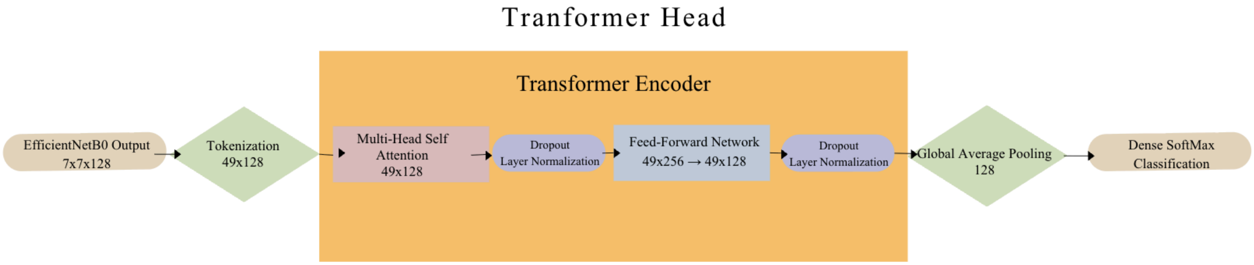

Tranformer Head
Transformer Encoder
EfficientNetB0 Output 7x7x128
Tokenization 49x128
Multi-Head Self Attention 49x128
Dropout Layer Normalization
Feed-Forward Network 49x256 → 49x128
Dropout Layer Normalization
Global Average Pooling 128
Dense SoftMax Classification

Fig. 2. The Transformer Head architecture which is attached onto the pretrained EfficientNetB0 CNN model. In the first training run, this is the only unfrozen and tuned portion of the model. In the second training run the top layers of EfficientNetB0 are unfrozen and tuned as well.

The output of EfficientNetB0 is a 7x7x1,280 feature map, representing each of the 7x7 image patches. The first stage of the transformer tokenizes this feature map into 49 vectors with 1,280 dimensions to prepare for self attention weighting, followed by a linear projection to 128 dimensions to reduce overfitting and improve runtime efficiency. The transformer encoder then applies Multi-Head Self Attention with 4 heads and a Dropout rate of 0.1, allowing each vector to “look at” every other vector, learning spatial dependencies. Residual Connections merge the inputs from the CNN Backbone with the attention spatial dependencies and Layer Normalization smooths those vectors for the Feed-Forward Network (FFN). The FFN consists of two dense layers, to learn non-linear combinations of local features. The first dense layer with Re-Lu activation expands to 49x256 tensors with added feature representation and passes into the second layer returning to the 49x128 token embedding dimension for Residual Connection. Residual Connections and another Layer Normalization complete the Transformer Encoder.

Following the Transformer Block, Global Average Pooling averages across the 49 vectors to produce a 128 dimension summary of the image. A Dropout rate of 0.35, which was found through tuning, is applied to reduce overfitting with the small sample size before our dense softmax classifier applies probabilities to each out the three classes.

Like the Transfer Learning model, a second fine tuning phase was run to fine tune the architecture and improve performance. Instead of using the Learning Rate Scheduler, a constant low learning rate ($1e^{-6}$) was set. The top 100 layers of EfficientNetB0 were unfrozen, reducing the additionally trained layers to a small subset of the pretrained model. This allowed for further training of the backbone CNN architecture to relate to the results of our Transformer Head addition.

f. Explainability Techniques

To assess the interpretability of the transfer learning model and examine how it processes chest X-ray inputs, three complementary explainability techniques were applied: Grad-CAM, gradient-based saliency, and occlusion sensitivity. Grad-CAM was used to generate coarse activation maps that highlight the spatial regions within an image that most influence the model’s classification decision. Gradient-based saliency maps provided finer pixel-level insight by visualizing how small changes in individual pixels affect the model’s output. Occlusion sensitivity offered a perturbation-based perspective by systematically masking small image patches and observing changes in predicted class confidence, thereby identifying regions that contribute most strongly to the model's internal representation. Together, these methods offer a multi-scale view of the model's feature utilization and help verify that its predictions align with meaningful image structures.

While standard Grad-CAM mapping is not possible for Transformers as they do not use 2D spatial operations like Convolution Layers, added interpretability from each token is still available through extracting token importance. Two methods are instituted to display this, token heat mapping to focus on each token's importance and a heat map overlay to relate contextually back to regions of the lung. By applying GradientTape, the model "watches" each token as it passes through the layers and calculates importance with the formula below

$$TokenImportance_i = \left(\frac{1}{128}\right) \sum_{j=1}^{128} |(\frac{\partial S}{\partial T_{i,j}})T_{i,j})|$$

where *i* represents the token index (49 tokens), *j* represents the feature index (128 features), $T_{i,j}$ is the activation value for a particular token/feature combination, and *S* is the class score logit for the predicted class. This formula uses the GradientTape's mapping of how sensitive *S* is to each feature activation, and multiplies that by the features activation value, taking the absolute value before averaging across all the features in that token, producing feature importance for that token. The importances are then normalized to a 0-1 scale and applied to a heat map to clearly describe each token individually. Then the importance values are upsampled to a 224x224 matrix to be overlaid on the original image, displaying the important regions of the lung in classification. This process allows for importance to include global interactions between regions as opposed to just local activations.

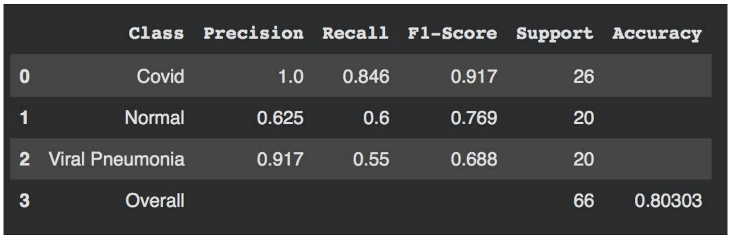
g. Evaluation Metrics

Model performance was assessed using accuracy, precision, recall, and F1-score, with class-specific metrics computed from the test set using scikit-learn. These metrics were chosen to provide a comprehensive view of both overall classification performance and sensitivity to each diagnostic category. In particular, recall was emphasized for the Covid class due to its importance in minimizing false negatives in a clinical screening context.

IV. Results & Discussion

a. Baseline Model Performance

The baseline CNN trained from scratch achieved a test accuracy of 80.30%, establishing the lower bound benchmark for the study. As shown in Table 2, the models reached strong performance on the Covid class with a precision of 1.000, recall of 0.846, and F1 score of 0.917. Performance decreased for the Normal and Viral pneumonia classes, with recalls of 0.6000 and 0.550, respectively, indicating difficulty in separating these two categories. The confusion matrix in Fig. 3 shows that most misclassifications occurred between Normal and Viral Pneumonia, with the model often confusing lower-contrast opacities. These results reflect the limitations of learning radiographic patterns solely from the training data without pretrained feature hierarchies.



	Class	Precision	Recall	F1-Score	Support	Accuracy
0	Covid	1.0	0.846	0.917	26	
1	Normal	0.625	0.6	0.769	20	
2	Viral Pneumonia	0.917	0.55	0.688	20	
3	Overall				66	0.80303

Table 2. This table reports class-level precision, recall, F1-score, and support for the baseline CNN trained from scratch. The model achieved an overall test accuracy of 80.30%. Class-specific metrics highlight performance variability across Covid, Normal, and Viral Pneumonia categories, providing a quantitative benchmark for comparison with the transfer learning and fine-tuned models.

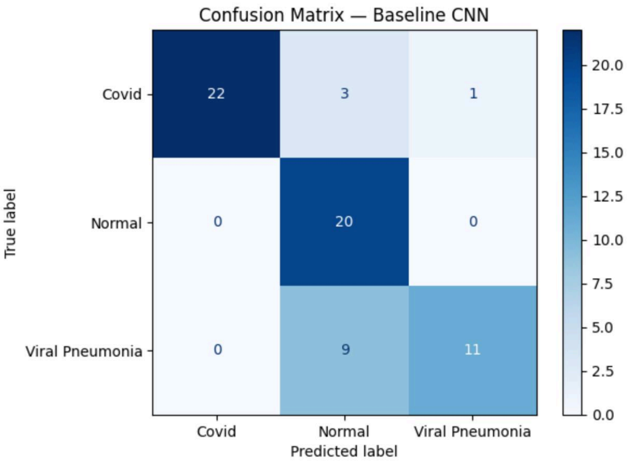
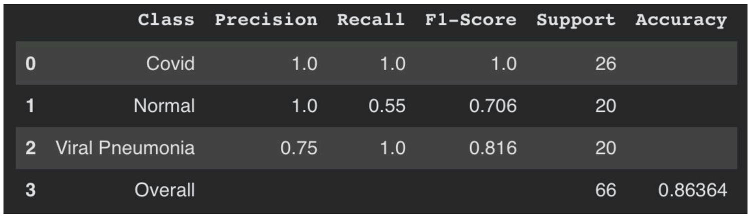


Fig 3. Confusion matrix for the baseline CNN model on the test set, illustrating per-class prediction performance across Covid, Normal, and Viral Pneumonia categories.

b. Transfer Learning Model Performance

The EfficientNetB0 transfer learning model achieved a higher overall test accuracy of 86.36%, outperforming the baseline by roughly six percentage points. This model demonstrated substantial improvements in class-specific metrics, particularly for Covid and Viral Pneumonia. The Covid class achieved perfect performance: precision 1.000, recall 1.000, and F1-score 1.000. Viral Pneumonia also improved dramatically, with recall rising from 0.550 to 1.000 and F1-score increasing to 0.816, all summarized in Table 3. The confusion matrix in Fig. 4 confirms that the transfer model correctly classified all Covid and Viral Pneumonia images. Normal-class performance remained more challenging, with a precision of 1.000 but recall of 0.550, suggesting partial overlap in radiographic appearance with the pneumonia category. Overall, the transfer learning model demonstrated stronger generalization and substantially better sensitivity to clinically significant findings.



	Class	Precision	Recall	F1-Score	Support	Accuracy
0	Covid	1.0	1.0	1.0	26	
1	Normal	1.0	0.55	0.706	20	
2	Viral Pneumonia	0.75	1.0	0.816	20	
3	Overall				66	0.86364

Table 3. Evaluation metrics for the EfficientNetB0 transfer learning model. The model achieves perfect precision, recall, and F1-score for the Covid class, along with notable improvements in Viral Pneumonia classification compared to the baseline model.

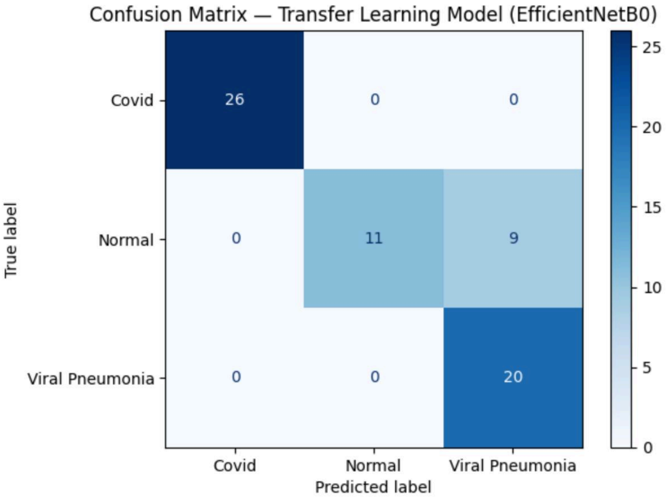
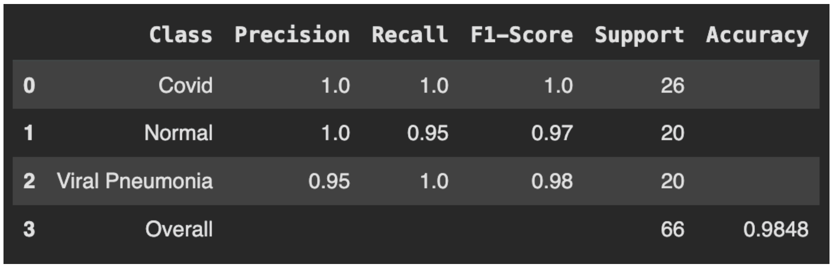


Fig 4. Confusion matrix for the EfficientNetB0 transfer learning model on the test set, illustrating per-class prediction performance across Covid, Normal, and Viral Pneumonia categories.

c. CNN + Transformer Model Performance

The addition of the Transformer Head to our model greatly improved our results, achieving an accuracy of 98.48%. As shown in Table 4. This model continued its perfect performance with Covid classification, with recall, precision, and F1-score being 1.000. The model also remains perfect at classifying those who truly have Pneumonia, with Pneumonia achieving a perfect recall of 1.00 as did the previous model, with precision also increasing to 0.95 and F1-Score increasing to 0.98. The classification for those with normal lungs also improved, as seen in Fig. 5 where only one test case was misclassified by this model, a normal lung being classified as Pneumonia. This gave Normal lungs precision of 1.00 recall of 0.95, and F1-Score of 0.97, the biggest jumps across all three categories from the Transfer Learning model. This suggests that the addition of global interaction between portions of the images provided much more clear separation between normal and Pneumonia lungs, an area that local representation struggled heavier with.



	Class	Precision	Recall	F1-Score	Support	Accuracy
0	Covid	1.0	1.0	1.0	26	
1	Normal	1.0	0.95	0.97	20	
2	Viral Pneumonia	0.95	1.0	0.98	20	
3	Overall				66	0.9848

Table 4. This table reports evaluation metrics for the CNN + Transformer model. The model achieves perfect precision, recall, and F1-score for the Covid class, perfect recall in the Viral Pneumonia class, and heavy improvements in both the Normal and Viral Pneumonia classes.

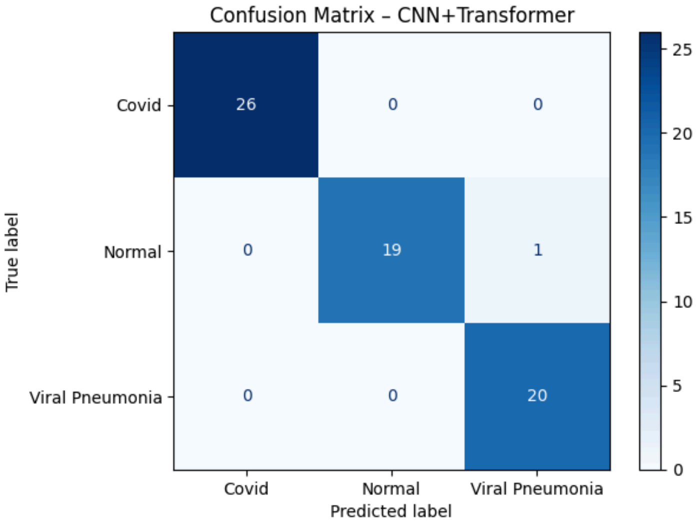
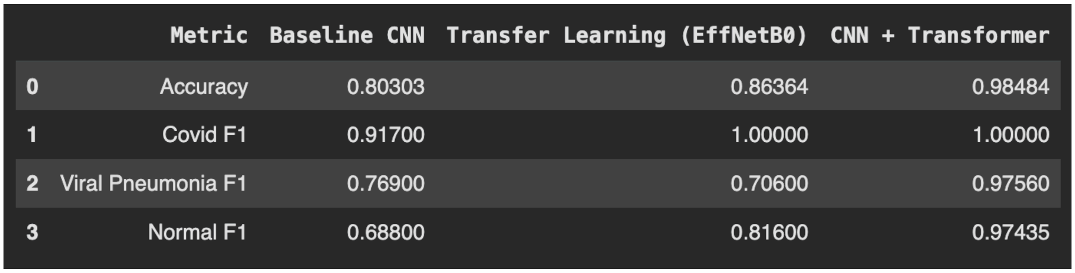


Fig 5. Confusion matrix for the CNN + Transformer model on the test set, illustrating per-class prediction performance across Covid, Normal, and Viral Pneumonia categories.

d. Model Comparison

A direct comparison of the baseline CNN, Transfer Learning, and Transformer models is summarized in Table 5. Macro-averaged metrics improved across the board when using EfficientNetB0: precision increased from 0.847 to 0.897, recall from 0.799 to 0.850, and F1-score from 0.791 to 0.842. These improvements reflect the benefit of pretrained hierarchy features, particularly for subtle opacity and texture patterns common in Covid and Viral Pneumonia cases. The transfer learning model also demonstrated more stable training, faster convergence, and fewer fluctuations in validation accuracy, which is typical when initializing from ImageNet weights.

When adding on the Transformer head, test results improved further, with near perfect results. Accuracy increased to 0.984, while Viral Pneumonia F1-score increased about 0.27 points to 0.976 and Normal F1-score increased about 0.158 points, up to 0.974. While the Transfer Learning model already perfectly classified Covid cases, the addition of global interactions in the Transformer model displayed much better separability between Normal lungs and Pneumonia lungs, with less false reports of Viral Pneumonia in healthy subjects. The model did take longer to reach stable convergence and longer run time overall, but after tuning achieved similar fluctuation in validation accuracy to the Transfer Learning model.



	Metric	Baseline CNN	Transfer Learning (EffNetB0)	CNN + Transformer
0	Accuracy	0.80303	0.86364	0.98484
1	Covid F1	0.91700	1.00000	1.00000
2	Viral Pneumonia F1	0.76900	0.70600	0.97560
3	Normal F1	0.68800	0.81600	0.97435

Table 5. Comparison of baseline CNN, EfficientNetB0 transfer learning, and CNN + Transformer performance across accuracy and class-specific F1-scores. Adding both Transfer Learning and Transformers improved results from the previous model, achieving perfect Covid classification and perfect classification of those with Viral Pneumonia with both methods applied.

e. Explainability Results

To better understand how the transfer learning model arrived at its predictions and to evaluate whether its decisions aligned with clinically meaningful patterns, three complementary explainability methods were applied: Grad-CAM, smoothed gradient-based saliency, and occlusion sensitivity. Fig. 6 illustrates one correctly classified example from each class, Covid, Normal, and Viral Pneumonia, demonstrating how the model's internal representations differ depending on the underlying pathology. These methods provide insight at multiple spatial scales: Grad-CAM reveals coarse regional activations within the lungs, saliency identifies fine-grained pixel-level contributions, and occlusion maps quantify how sensitive the model is to local perturbations in specific anatomical regions.

Across the Covid examples, Grad-CAM consistently highlighted bilateral lower-lung regions, which often correspond to ground-glass opacities characteristic of Covid-19 infection. The overlays show dominant activation in the mid-to-lower lung zones, suggesting the model has learned clinically relevant discriminative features. Saliency maps further reinforce this interpretation by revealing dense, high-intensity gradient activity concentrated within similar regions. Occlusion maps for Covid demonstrate that masking these areas causes substantial drops in model confidence, confirming their importance in the decision-making process.

The Normal example exhibited the opposite behavior. Grad-CAM maps showed minimal activation within the lung fields, with scattered, low-intensity regions instead of focused patterns. The saliency map displayed diffuse low-level gradients, reflecting the model's recognition that no dominant pathological structures were present. The occlusion map similarly showed weak sensitivity across most regions, indicating that masking patches of a Normal X-ray does not substantially alter the predicted probability. This behavior demonstrates that the model is not overly sensitive to irrelevant structures and appropriately differentiates "absence of findings" cases.

For Viral Pneumonia, the explainability outputs displayed distinct activation patterns compared to Covid. Grad-CAM highlighted more asymmetric, localized regions of interest, consistent with the patchy or unilateral infiltrates common in viral pneumonia. Saliency maps exhibited finer localized texture sensitivity in these regions, and occlusion maps showed pronounced confidence drops when these asymmetric areas were masked. These findings indicate that the model does not simply rely on a generalized "pneumonia-like" pattern but instead identifies class-specific signatures differentiating Viral Pneumonia from Covid and Normal images.

Taken together, these explainability results demonstrate that the model's predictions rely on meaningful radiographic features and vary appropriately across diagnostic categories. The agreement between Grad-CAM, saliency, and occlusion maps provides convergent evidence that the model's internal representations align with clinically interpretable lung regions, lending confidence to its behavior and supporting its reliability as a diagnostic aid.

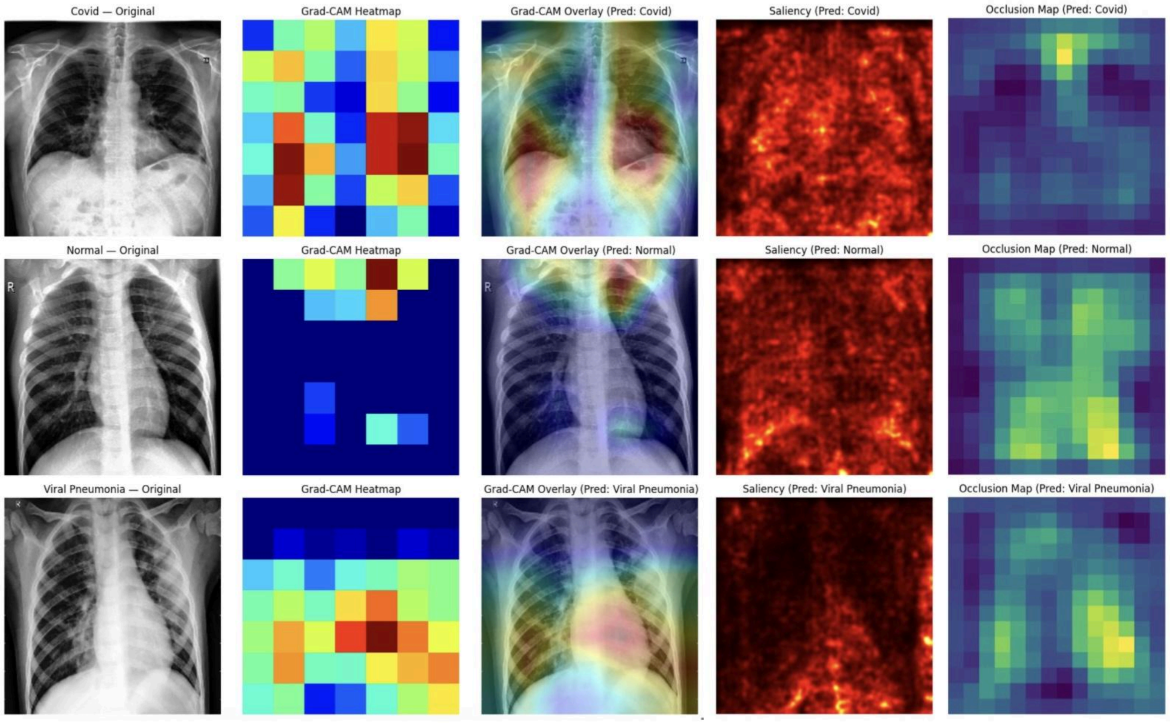

Covid — Original
Grad-CAM Heatmap
Grad-CAM Overlay (Pred: Covid)
Saliency (Pred: Covid)
Occlusion Map (Pred: Covid)
Normal — Original
Grad-CAM Heatmap
Grad-CAM Overlay (Pred: Normal)
Saliency (Pred: Normal)
Occlusion Map (Pred: Normal)
Viral Pneumonia — Original
Grad-CAM Heatmap
Grad-CAM Overlay (Pred: Viral Pneumonia)
Saliency (Pred: Viral Pneumonia)
Occlusion Map (Pred: Viral Pneumonia)

Fig 6. Explainability visualizations for Transfer Learning model with one correctly predicted example from each class (Covid, Normal, and Viral Pneumonia). Columns show, from left to right: the original chest X-ray, Grad-CAM heatmap, Grad-CAM overlay, smoothed gradient-based saliency map, and occlusion sensitivity map. Grad-CAM highlights high-level spatial regions most influential to the predicted class, saliency reveals pixel-level gradients contributing to the model decision, and occlusion maps measure the change in class confidence when local image regions are masked. These combined visualizations provide a multi-scale assessment of how the transfer learning model interprets radiographic features for each diagnostic category.

In the final Transformer head model, two more explainability methods were applied to add another layer of information, this time tracking activation signals throughout the lung X-Ray images which include not just local activations, but also signals of global interactions between various regions of the lung. While standard Grad-CAM processing requires extraction from a convolution layer, the tracking of token importance with the use of GradientTape allows for similar style visualizations which focus on the activations and signals tracked through each of the 49 token patches of the images. For visualization purposes, token heatmaps were generated for easy comparison and understanding to the Grad-CAM results, along with token importance overlays to relate the findings back to contextually relevant information related to lung regions, which are displayed in Fig. 7. In addition, softmax probabilities were plotted to add insights into how the model distinguished between classes. These visualizations were gathered for example images of each class for cross-comparison.

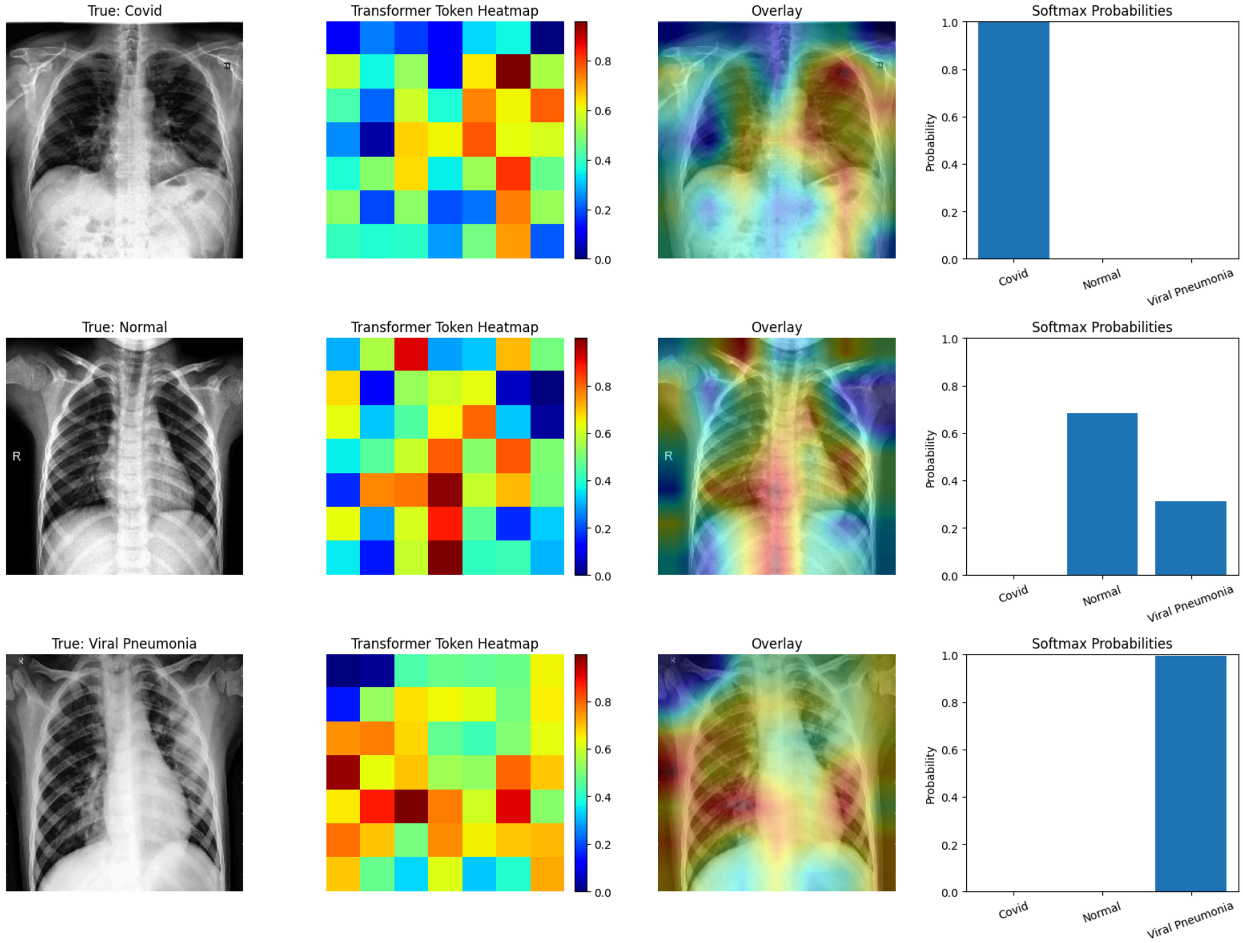


Fig 7. Explainability visualizations from the CNN + Transformer model with one correctly predicted example from each class (Covid, Normal, and Viral Pneumonia). Columns show, from left to right: the original chest X-ray, token heatmap, token heatmap overlay, and softmax probabilities. Token heatmaps and overlay relay similar information to Grad-CAM visualizations, adding global interactions between lung regions in importance calculations. Softmax probabilities distinguish derived probabilities of each class from the model for each image example. These combined visualizations provide a multi-scale assessment of how the CNN + Transformer model interprets radiographic features for each diagnostic category.

Overall, the findings from the Transformer model explainability show enhanced representation of what was seen in the Grad-CAM extractions. The COVID cases still show signals of bilateral activations, although there is much more emphasized importance in the left lung (right lung in the X-Ray image). The activations also show more importance throughout the entire lungs as opposed to being isolated to the bottom portion, particularly showing the strongest importance in the upper portion of the left lung. The importance overlay compared to the Grad-CAM reveals that difficult to find local activations in the upper region of the lung actually display very high importance when they are interacting with signals in the lower part of the lung, showing the importance of global understanding of these images and its ability to distinguish COVID, as it achieved 100% confidence on this image in its COVID classification.

The Normal lung showed very similar behavior to the Grad-CAM findings, in that minimal activation was found within the lung fields. The overlay displays that by far the tokens of greatest importance were found to be along the spinal cord, a very interesting distinction for the model. By having the multi-class groupings, the model was able to derive the two main weakness areas through the other illness, and then distinguish the Normal lungs, likely by the lack of activations in the tokens represented by the lungs themselves. This is a great discovery, as the model was essentially able to learn on its own what the true "baseline" X-Ray of a lung should look like, allowing for accurate distinction of the illness groups through lung activations themselves. It is important to note though, and as has been reflected in both the CNN + Transformer and Transfer Learning model results that our model was much less confident in its normal lung predictions, as the activations that do show in the lungs in the lower areas of each side do resemble lesser importance of what is noticed in the Viral Pneumonia lungs, leading to split probabilities between Normal and Pneumonia classifications, along with the misclassification noted in Fig. 5.

The Viral Pneumonia overlays once again support the findings of the Grad-CAM extraction, showing the regions of heaviest importance being isolated to the lower portions of the lungs. A noticeable flip between the lung side from the Grad-CAM images to the token heatmaps does occur, shifting from the left to the right lung respectively. This displays more the interaction between both sides of the lower regions of the lungs, highlighting the overall importance and medical context of locating the lower regions of the lungs for Viral Pneumonia. The importance of isolating to the lower region of the lungs is also emphasized by the difference between the Normal and Pneumonia overlays. While both show importance in similar areas of the X-Ray, the Pneumonia shows its highest importance in the lung regions themselves, while Normal focuses on the spinal cord more heavily. When the Normal classification showed doubt in its soft-max probabilities, once the lower lung regions were able to reach a higher threshold of overall importance, the model was able to classify Pneumonia with 100% confidence.

Overall, the results from the CNN + Transformer model explainability methods further support the use of deep learning methods on X-ray imaging to both diagnose COVID-19 and Viral Pneumonia and understand what regions of the lungs are affected in each case. The showing of enhanced learning is evident in the jump from the Grad-CAM extractions from the Transfer Learning model to the Token Importance mapping of the Transformer model through heavier emphasis of similar regions between the two. The awakening of regions across the lungs as well also display the importance of including global interactions in the diagnostic model, available through the addition of the Transformer head, that may go unnoticed by the Transfer Learning model alone.

f. Summary of Findings

The experimental results demonstrate that the EfficientNetB0 transfer learning model substantially outperforms the baseline CNN in both global and class-specific metrics. The model not only achieved a higher overall accuracy but also delivered perfect sensitivity and precision for the Covid class. The explainability analyses further show that the model's predictions are grounded in reasonable anatomical regions, indicating alignment with known radiographic patterns of Covid infection.

When adding a Transformer head to the Transfer Learning model, results greatly improved again, achieving near perfect results with only one misclassification on its best run. Reaching an accuracy of over 98%, the model was also able to perfectly classify those who are truly ill with COVID-19 and Pneumonia from those who have healthy lungs, while also providing further distinction in the Normal class. This displays high importance in the application of Transformers in image recognition tasks for lung X-Rays, and also points towards overall importance of this method in health imaging deep learning tasks in general. The ability to not only distinguish local signals in images, but also global interactions across lung regions proved to be very insightful for the model's predictive power, allowing the model to more closely relate to true lung anatomical function. The explainability methods assessed also supported the use of the transformer models in medical settings, as the results were able to directly relate back to anatomical regions of the lungs of interest for the illness cases, and even going as far as lowering activations in lung regions for healthy lungs, correctly identifying the baseline. This provides more validity to the models ability to understand the context of the images and be able to locate similar patterns in other lung X-rays.

One limitation was the limited size of the dataset available, which affected the stability of our results. While the reported results represent the abilities of the model and generally how the models perform, the results from differing runs are not always consistent, seemingly at some mercy of the seed being set. Methods were applied to achieve more consistent results, and these improved stability, including Learning Rate Scheduler, secondary rounds with enhanced fine tuning, early stopping, and added epochs. Still some variance in results persists, and would be best remedied by being able to test on a larger dataset. Another limitation is general variability in X-ray techniques and practices across medical institutions. While this issue was addressed through augmentations of the images within reasonable ranges, representative of these variations, in a medical setting where precision is absolutely necessary these models should be tested for generalizability on images from a wider variety of sources before clinical approval is possible.

V. Conclusions and Future Work

This study demonstrated that transfer learning with EfficientNetB0 provides meaningful improvements over a baseline CNN for Covid-19 chest X-ray classification. While the baseline model established a reasonable benchmark with an accuracy of 80.30%, the transfer learning model achieved higher overall accuracy and perfect sensitivity for the Covid class, indicating a stronger ability to detect clinically significant cases. The multi-scale explainability analyses, including Grad-CAM, saliency maps, and occlusion sensitivity, confirmed that the model relied on physiologically relevant lunch regions rather than artifacts, supporting the interpretability and trustworthiness of its predictions. Together, these findings highlight the effectiveness of pretrained convolutional architectures when applied to limited medical imaging datasets.

The study also demonstrates how continuous improvement can be achieved through the addition of a Transformer, particularly through the Transformer's tokenization process which allows for global interactions between lung regions. This provides direct insight into the question of this paper, segmenting COVID-19, Viral Pneumonia, and Normal lungs for diagnosis, but also provides insight into the best methods for implementing Deep Learning methods for medical imaging tasks in general. Human anatomy is a complex network with many nodes and regions working in tandem. The great improvement in model results, achieving only one misclassification, going from local feature extraction to global shows how Transformers provide extreme improvement in how Deep Learning models can learn the full context of how the body is connected on a more detailed level, and how fine interactions within the body can be understood in a way that is blind to the naked eye or simpler models.

In the context of medical diagnosis with Deep Learning and artificial intelligence techniques, sensitivity of illness classes plays the leading factor. It is far less consequential to misdiagnose a healthy individual as having an illness and providing extra monitoring and treatment to that individual than to misdiagnose an individual who is ill as being healthy, particularly when dealing with potentially severe lung complications such as Viral Pneumonia or Covid. Both the Transfer Learning and CNN + Transformer models were able to achieve perfect performance in this regard, and the CNN + Transformer also showed improved classification for the Normal lungs, which is still beneficial in lowering patient costs and stress for those who are truly healthy.

While limited by a small dataset, and the requirements for extensive clinical testing before allowing implementation of Deep Learning models for clinical diagnosis use, this model shows very strong indications that these methods are certainly applicable and can lead to an improvement in early diagnosis of these complications, likely preventing serious illness and death in many cases. They also display the ability to fill the gaps that are missed by the current standard RT-PCR tests, which only detect COVID-19 in a short and specific time range since exposure and do not provide information towards the potential of developing more serious complications in the lungs.

Future work may focus on expanding the dataset to include more diverse and higher-quality radiographs, which could improve generalization and stabilize class-specific performance. Additional architectural explorations, such as deeper EfficientNet variants or vision transformers (ViTs), may further enhance discriminative power. It is important to note that deeper networks and especially ViTs will most likely require a much larger dataset to achieve similar or better results than these models achieved. Integrating segmentation methods to isolate lung fields prior to classification could also reduce noise from irrelevant anatomy. Finally, ensembling multiple pretrained models or incorporating clinical metadata may yield further performance gains and produce a more robust, clinically useful diagnostic tool.

In addition to improving the models, an implemented system where doctors can directly upload an X-ray image and process the results in real time to achieve a classification would be monumental in applying models to clinical settings. This would save doctors and patients a lot of time and hospitals a lot of money which can be reapplied to improved treatment of conditions. A call to action is to improve the shared access to X-ray images for this type of classification to allow for greater research into this topic. Similar repositories are already implemented in other fields, for example the Autism Brain Imaging Data

Exchange (ABIDE)[6], which is a shared repository of brain MRI scans across many hospitals and research centers, which has greatly enhanced research into autism diagnosis. This study shows incredible potential towards applying deep learning to diagnosing these ailments, and the ability to share medical imaging will allow continuous improvements of these models to a point of clinical implementation. In the case of this study, this would greatly improve the ability to achieve more stable and consistent results which can be clinically accepted.

VI. References

Akter S, Shamrat FMJM, Chakraborty S, Karim A, Azam S. COVID-19 Detection Using Deep Learning Algorithms on Chest X-ray Images. Biology (Basel). 2021 Nov 13;10(11):1174. doi: 10.3390/biology10111174. PMID: 34827167; PMCID: PMC8614951.

Apostolopoulos ID, Mpesiana TA. Covid-19: automatic detection from X-ray images utilizing transfer learning with convolutional neural networks. Phys Eng Sci Med. 2020 Jun;43(2):635-640. doi: 10.1007/s13246-020-00865-4. Epub 2020 Apr 3. PMID: 32524445; PMCID: PMC7118364.\

Di Martino, A., Yan, C. G., Li, Q., Den Boer, J. A., Falini, A., Gordon, C., & Milham, M. P. (2014). The autism brain imaging data exchange: background, rationale, and implementation. *Scientific data*, *1*, 140010.

Hajar Lamouadene, Majid EL Kassaoui, Mourad El Yadari, Abdallah El Kenz, Abdelilah Benyoussef, Amine El Moutaouakil, Omar Mounkachi, Detection of COVID-19, lung opacity, and viral pneumonia via X-ray using machine learning and deep learning, Computers in Biology and Medicine, Volume 191, 2025, 110131, ISSN 0010-4825, https://doi.org/10.1016/j.compbiomed.2025.110131.

Islam R, Tarique M. Chest X-Ray Images to Differentiate COVID-19 from Pneumonia with Artificial Intelligence Techniques. Int J Biomed Imaging. 2022 Dec 22;2022:5318447. doi: 10.1155/2022/5318447. PMID: 36588667; PMCID: PMC9800093.

R. L. Rubin et al., "The Role of Chest Imaging in Patient Management During the COVID-19 Pandemic: A Multinational Consensus Statement from the Fleischner Society," Radiology, vol. 296, no. 1, pp. 172–180, Jul. 2020, doi: 10.1148/radiol.2020201365.

[6] Di Martino, A., Yan, C. G., Li, Q., Den Boer, J. A., Falini, A., Gordon, C., & Milham, M. P. (2014). The autism brain imaging data exchange: background, rationale, and implementation. *Scientific data*, *1*, 140010.