COVID-19 Detection from Chest X-Ray Images Using AI Models

Data Mining Course Term Project

Han-Sheng Huang

National Taiwan University

Contents

1 Background

Since the outbreak of COVID-19, the pandemic has caused over three million deaths worldwide and infected countless individuals, making the control of virus spread an urgent issue for governments and public health experts. To prevent large-scale transmission, most countries currently rely on testing and isolating confirmed cases.

Therefore, accurately and rapidly identifying infected individuals is crucial. The standard screening method, nucleic acid testing (Reverse Transcriptase Polymerase Chain Reaction, or RT-PCR), is costly, time-consuming, and requires specialized equipment and trained medical personnel. Using AI models to detect COVID-19 infection could provide faster results with high accuracy while reducing manpower requirements. This approach would enable efficient and rapid identification of infected patients, allowing timely isolation and helping prevent widespread outbreaks.

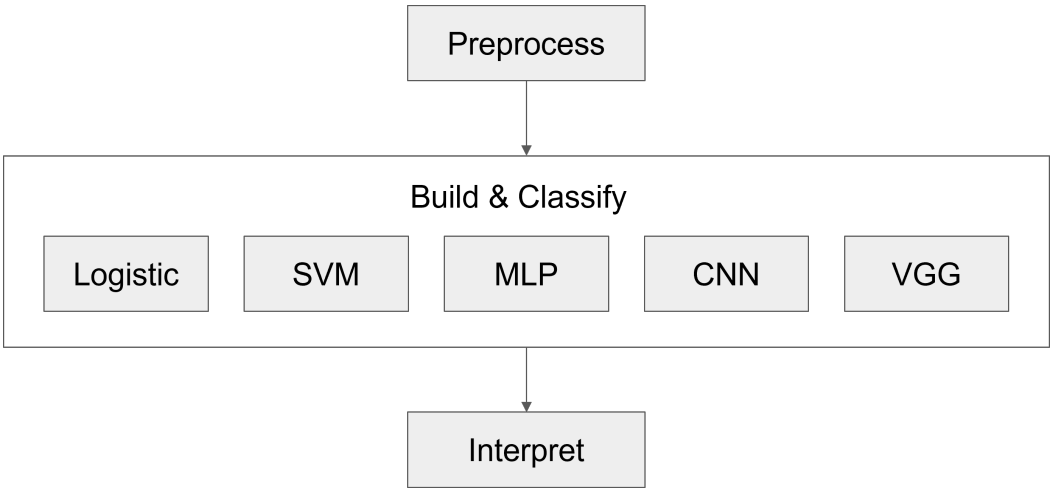
2 Research Framework

Since COVID-19 primarily affects lung function, examining a person's lung condition may help determine whether they are infected. Building on previous research, this report aims to use an AI model to analyze chest X-ray images and identify whether an individual has COVID-19.

1. Preprocess
   Clean and preprocess chest X-ray data, apply data augmentation, and use transfer learning techniques to address the limited dataset.

2. Build Model + Classify
   Develop multiple classification models—including Logistic Regression, SVM, MLP, CNN, and VGG—to determine whether an individual is infected with COVID-19 based on chest X-ray images.

3. Interpret
   Apply LIME to interpret model predictions and highlight important regions of the X-ray images, providing medical professionals with insights into the model's decision-making process.



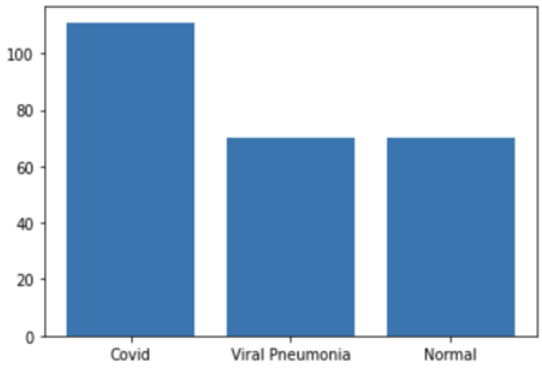
3 Data Collection

The dataset used in this study was obtained from the Kaggle website's "COVID-19 Image Dataset" (`https://www.kaggle.com/pranavraikokte/COVID19-image-dataset`). This dataset consists of X-ray images, which are further categorized into three classes: COVID, Normal, and Viral Pneumonia.
Specifically, the COVID category contains chest X-rays of patients infected with COVID-19; the Normal category represents chest X-rays of healthy individuals; and the Viral Pneumonia category includes X-rays of patients infected with common viral pneumonia. The figure below shows examples from each class:

Regarding the dataset size, the provider has already divided the data into a training set and a test set. The training set contains a total of 251 images (111 COVID, 70 Normal, and 70 Viral Pneumonia), while the test set contains 66 images (26 COVID, 20 Normal, and 20 Viral Pneumonia).

4 Data Preprocessing

4.1 Data Imbalance



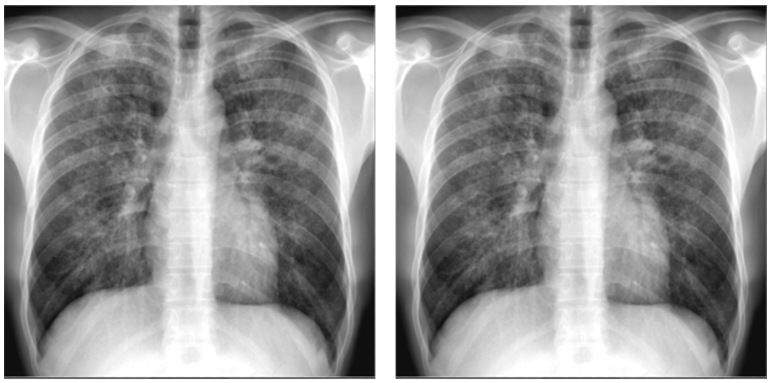
The figure above illustrates the number of images in each class. As shown, the original training dataset exhibits a slight class imbalance (111, 70, 70). To prevent this imbalance from biasing model training, Keras' ImageDataGenerator was employed to generate additional images. Augmentation parameters included rotation and scaling. Using the class with the largest number of images (COVID) as the reference, the other two classes were balanced to 111 images each.

4.2 Small Sample Size

The training dataset is relatively small, containing only 251 images. Even after balancing, the total number of images reaches only 333. For deeper models, such a limited sample size may lead to suboptimal training performance. Therefore, the dataset was further augmented to increase each class to 500 images.

4.3 ZCA Whitening

Initial observations of the data revealed differences in the grayscale (shadow) patterns among the X-ray images of different classes. COVID images tend to be relatively brighter, with the lower regions mostly white, while Normal images are darker with fewer shadows. It was hypothesized that applying ZCA whitening could not only reduce noise but also enhance the distinction between light and dark regions, potentially aiding classification. Therefore, ZCA whitening was applied, and the dataset was augmented to 500 images per class. Due to the high RAM requirements of ZCA whitening, the images were resized to 80×80 pixels during the whitening process in this training. As a result, the ZCA-processed images shown above have lower resolution.



Original Images After ZCA Whitening

4.4 Flattening to One-Dimensional Arrays

Since some models cannot be trained directly on image files, all datasets were converted into one-dimensional arrays of consistent size for model training, in addition to using the original images. ZCA-whitened images were flattened to 6,400 dimensions (80×80), while the remaining images were flattened to 65,536 dimensions (256×256).

4.5 Resulting Datasets After Preprocessing

After preprocessing, four datasets were generated:

1. Original Dataset: The original dataset.
2. Balanced Dataset (111): Each class balanced to 111 images.
3. Balanced Dataset (500): Each class balanced to 500 images.
4. ZCA (80×80): Images processed with ZCA whitening and each class balanced to 500 images.

5 Model Development

5.1 SVM

The purpose of Support Vector Machines (SVM) is to find the optimal hyperplane that maximizes the margin separating different classes. The margin is determined by the samples closest to the decision boundary, which are called support vectors, giving the model its name. During training, different kernel functions—such as linear, polynomial, or RBF—can be selected to suit the characteristics and requirements of the data classification task.

5.2 MLP (Multilayer Perceptron)

MLP is a type of Neural Network (NN). Neural networks perform nonlinear feature transformations through artificial neurons, assigning different weights (W) to a feature vector (X) and applying activation functions to handle nonlinear classification problems. An MLP uses multiple hidden layers to learn and capture complex, important features for classification, thereby achieving improved classification performance.

5.3 CNN (Convolutional Neural Network)

CNN is one of the methods in deep learning, commonly applied in image recognition and natural language processing (NLP). Its specialized architecture enables the model to capture important features of images, thereby achieving strong capabilities in image recognition and word understanding.
Convolution: Convolution involves extracting features from an image, identifying the most relevant ones, and ultimately using them for classification.
Pooling: Pooling reduces the dimensionality of the feature map while retaining important features. This process decreases the number of parameters and helps prevent overfitting.

5.4 VGG (Visual Geometry Group)

The VGG (Visual Geometry Group) model is a classic neural network architecture, commonly seen in variants such as VGG16 and VGG19, distinguished by the number of layers.
The VGG19 model used in this study consists of 16 convolutional layers and 3 fully connected layers. Unlike earlier CNN models, VGG employs smaller convolutional filters (3×3) to extract more detailed information from images, and multiple layers of non-linear activations increase the overall non-linearity of the model. For pooling layers, the designers suggested that smaller pooling sizes help retain more information, which is why a 2×2 pooling kernel is adopted in the VGG architecture. The figure below illustrates the VGG19 model.
In addition to adjustments in the convolutional and pooling layers, VGG also applies specific processing techniques for training and testing data. For the training set, VGG employs multiple scale training, where the shorter side of each image is randomly scaled to a size between 256 and 512, and then a central 224×224 region is cropped as input for network training. For the test set, multiple crop testing is used: images are first resized to 280×280, and 224×224 crops are taken from the four corners and the center for testing. The final prediction is obtained by averaging the Softmax outputs from all crops.
Regarding the advantages and disadvantages of VGG, its strength lies in the simplicity and clarity of its design. Each convolutional layer uses a 3×3 filter, and each pooling kernel has a 2×2 size. Experiments have also shown that deeper models outperform shallower ones. However, the increased depth results in a larger number of parameters, which significantly raises computational costs during training.

6 Model Performance

The following presents the training results of different models applied to each dataset.

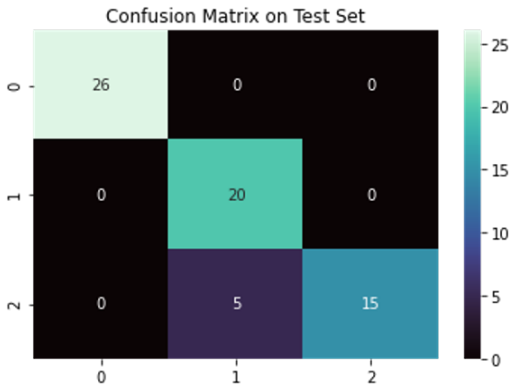
	Logistic Regression	SVM	MLP	CNN	VGG
Original Dataset					
Accuracy	84.80%	86.40%	81.80%	89.40%	92%
Precision	84.90%	86.00%	84.10%	90.30%	94%
Recall	83.70%	86.20%	81.50%	89.50%	92%
F1-score	0.83	0.86	0.81	0.89	0.92
Balanced Dataset (111)					
Accuracy	78.80%	90.90%	75.70%	92.80%	91%
Precision	78.20%	91.20%	63.90%	88.40%	93%
Recall	77.80%	90.80%	63.20%	86.50%	91%
F1-score	0.78	0.91	0.58	0.86	0.91
Balanced Dataset (500)					
Accuracy	84.80%	94.00%	83.30%	93.90%	98%
Precision	84.30%	94.40%	85.60%	93.70%	99%
Recall	84.10%	94.10%	83.20%	93.30%	98%
F1-score	0.84	0.94	0.83	0.93	0.98
ZCA (80x80)					
Accuracy	77.30%	70.00%	65.10%	68.20%	89%
Precision	76.50%	70.20%	65.20%	68.60%	92%
Recall	76.50%	69.40%	64.40%	67.30%	89%
F1-score	0.76	0.69	0.65	0.67	0.89

Based on the results in the table above, the following observations can be made:

1. Logistic Regression and SVM models achieved reasonably good results. VGG performed the best overall, possibly due to the use of pre-trained weights.

2. MLP showed the poorest performance, which may be attributed to suboptimal hyperparameter settings or model architecture. In comparison, CNN appears to be better suited for image recognition tasks.

3. From the confusion matrix, the COVID category consistently achieved the highest classification accuracy, while the other two categories were sometimes misclassified. For example, using the VGG model on the Original Dataset, class 0 corresponds to COVID, and classes 1 and 2 correspond to Normal and Viral Pneumonia, respectively. The classification results for the latter two categories are noticeably poorer.



4. ZCA preprocessing did not improve classification performance. This may be because X-ray images are less suitable for such preprocessing (as only the bones remain), and the reduction in image size (from 256 to 80) to save computation time may have caused information loss.
5. Training on the Balanced Dataset (500) generally yielded better results than the Original Dataset.
6. SVM outperformed the self-trained MLP, possibly because SVM is better suited for smaller datasets, whereas the MLP model's complexity in architecture and parameter selection likely contributed to its poorer performance.

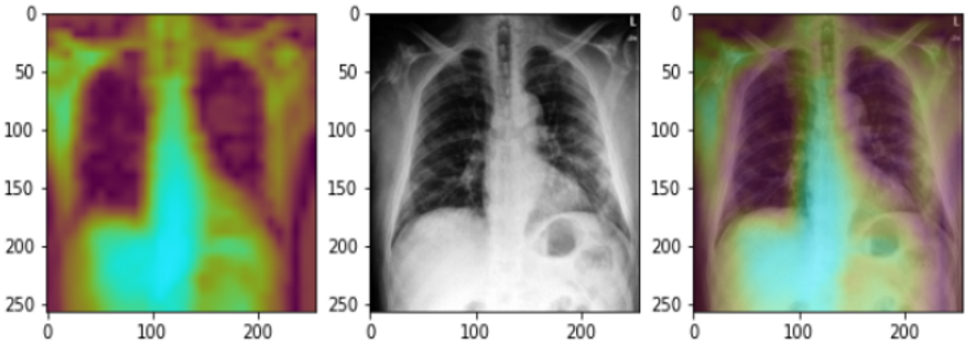
7 Model Interpretation

7.1 Grad-CAM (Gradient-weighted Class Activation Mapping)

Grad-CAM allows understanding of which regions of an image a CNN focuses on when performing image classification, thereby explaining the model's classification decision.
Class Activation Map (CAM) works by connecting a Global Average Pooling (GAP) layer after the last convolutional layer, instead of using the traditional Flatten Fully Connected Layer. After GAP transformation, each neuron corresponds to a specific feature map in the final layer. The weights connecting the GAP layer can be interpreted as the importance of each feature map for predicting the target class. By weighting each feature map according to its corresponding weight and summing them, the CAM (Class Activation Map) is obtained.
From the description of CAM, it is clear that implementing CAM requires connecting a GAP layer to the output of the last convolutional layer. However, this constraint restricts the flexibility of the network architecture. Grad-CAM addresses this issue by using backpropagation to compute the weights w used in CAM. As a result, Grad-CAM can generate activation maps regardless of the type of neural network layer used after the convolutional layers, enabling CAM without modifying the original model architecture.

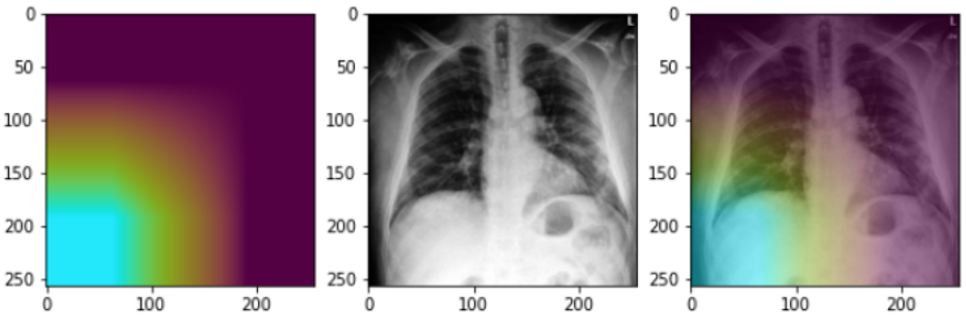
7.1.1 CNN Interpretation Results (using the 25th image from the COVID class as an example)

The following shows the output of the first convolutional layer:



From the Grad-CAM results, it can be observed that the important regions in the first convolutional layer of the CNN correspond mainly to the overall bone structure. The heatmap highlights areas such as the spine and the heart in the X-ray image, indicating that these regions are most significant for this layer.

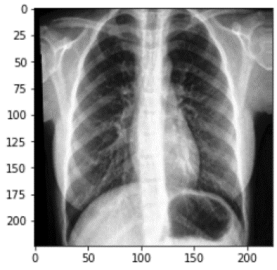
The following shows the output of the second convolutional layer:



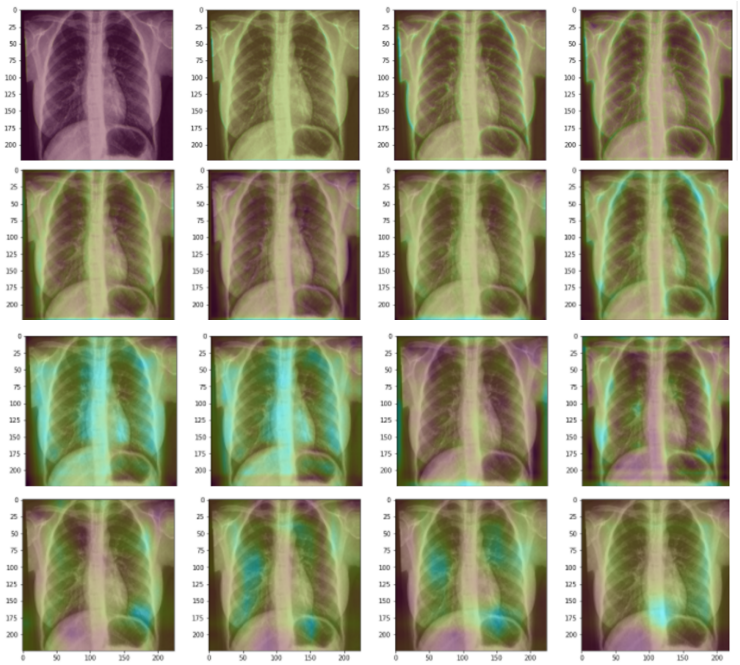
From the Grad-CAM results, it can be observed that the important regions in the second convolutional layer shift to the lower part of the left lung. Compared to the first layer, the highlighted features are more localized and detailed, suggesting that the lower left lung may exhibit specific abnormalities associated with COVID.

7.1.2 VGG-16 Interpretation Results (using the 40th image from the Normal class as an example)

The following shows the original image of the 40th Normal sample:



VGG-16 contains 16 convolutional layers, and the following are the heatmaps for each layer in sequential order:



In the earlier layers, the model primarily focuses on the overall structure and the skeletal regions. In the deeper layers, attention shifts to more specific areas of the lungs, such as the upper and lower sections of the left lung, the upper and lower sections of the right lung, as well as the spine and heart regions. By collaborating with medical professionals who possess domain expertise, these highlighted areas can be more precisely interpreted, aiding in infection detection and supporting subsequent clinical treatment decisions.

7.2 LIME (Local Interpretable Model-Agnostic Explanations)

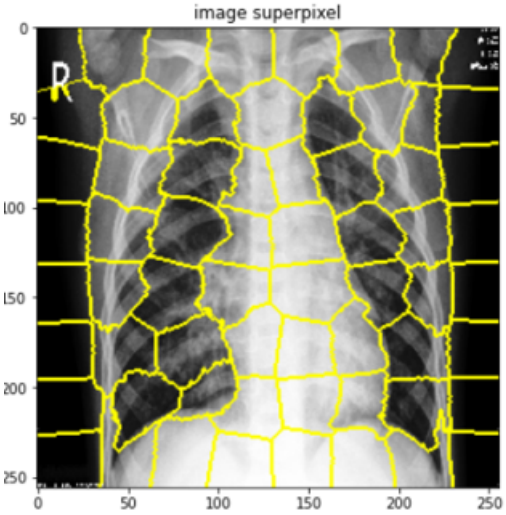
In addition to the relatively coarse heatmaps provided by Grad-CAM, LIME was also applied to investigate which specific regions of an image are important for different classes.

Before applying LIME, pixels with similar characteristics are grouped into larger segments called superpixels. In this work, SLIC (Simple Linear Iterative Clustering) was used to generate superpixels. SLIC employs an algorithm similar to K-means, but restricts distance calculations to a limited local region, reducing computational cost while producing more regular and coherent segmentations.

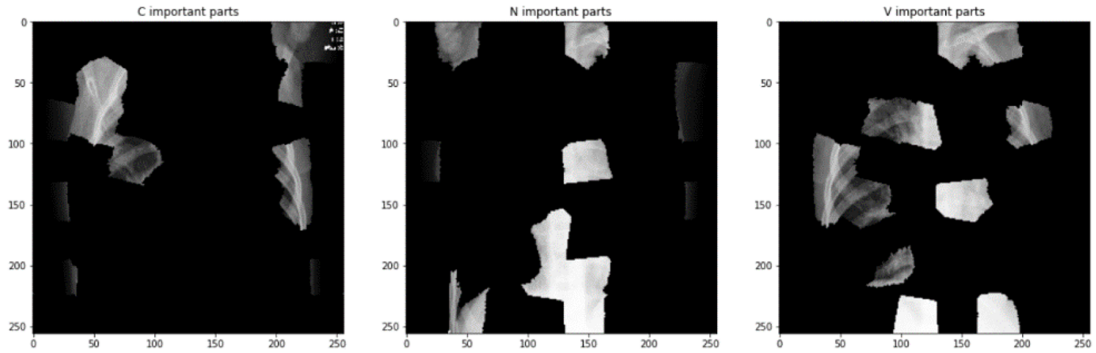
LIME then creates multiple perturbed image samples by randomly turning these superpixels on or off, generating different combinations of superpixel configurations. These perturbed samples are fed into the pre-trained complex model to obtain predicted probabilities for each class. Finally, a simple linear regression model is fitted, where X represents the presence or absence (1/0) of each superpixel configuration, and Y represents the predicted probability of a specific class. Through this linear model, the superpixels with larger weights can be identified—indicating that their presence has a significant impact on the model's predicted probability—thus revealing the regions most important for that class.

7.2.1 CNN Interpretation Results (using the 150th image from the COVID class as an example)

Superpixel segmentation of the image:



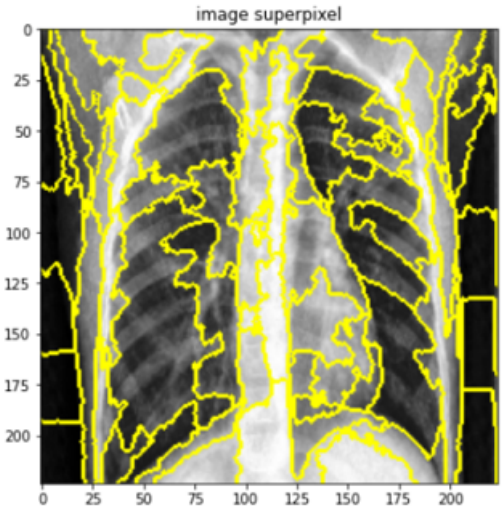
Important superpixels for different classes:



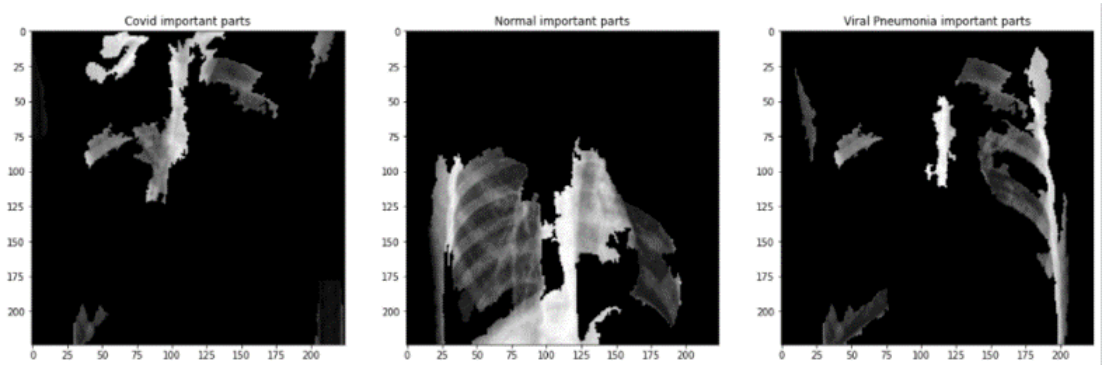
For the COVID class, the important superpixels are primarily located in the upper regions of the lungs, indicating that these areas have a strong influence on predicting COVID. For the Normal class, the key superpixels correspond to the central spine and heart regions, while for Viral Pneumonia, the important regions are mainly in the left lung.

7.2.2 VGG-16 Interpretation Results (using the 25th image from the Normal class as an example)

Superpixel segmentation of the image:



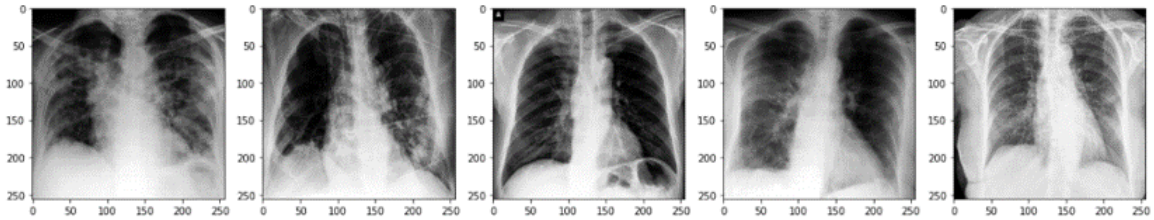
Important superpixels for different classes:



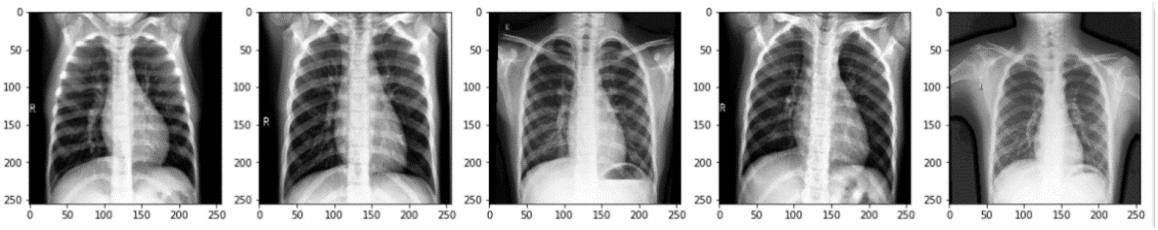
For the COVID class, the important superpixels are primarily located in the upper regions of the lungs, indicating that these areas have a strong influence on predicting COVID. For the Normal class, key superpixels correspond to the central spine, heart, and the lower part of the left lung, while for Viral Pneumonia, the important regions are mainly in the upper part of the right lung.

7.3 Other Interpretations

The original images were also examined to analyze the model's results. The following shows a lung X-ray image of a patient infected with COVID:



The following shows an X-ray image of a healthy lung:



From the two sets of images above, it can be observed that, compared to healthy lungs, the lungs of infected patients exhibit unclear regions around the heart and spine due to pulmonary abnormalities. The ribs also appear less distinct, and noticeable white patches are present. Some areas of the lungs show signs of atrophy. These characteristics may explain why the model performs so well and why it identifies specific regions as particularly important for classification.

8 Conclusion

In this study, logistic regression, SVM, MLP, a custom-designed CNN, and transfer learning using VGG were applied to predict categories from X-ray images, including COVID, Viral Pneumonia, and Normal.

For data preprocessing, due to the limited size of the original dataset (only 251 training samples and 66 testing samples), three additional datasets were generated using rotation and scaling for prediction: Balanced Dataset (111), Balanced Dataset (500), and ZCA(80×80) after ZCA whitening.

Through the analysis and methods described above, the model can detect infection with a high probability (up to 98%) and identify which specific regions of the lungs are most important, increasing model interpretability and trustworthiness. Furthermore, data augmentation via rotation and scaling effectively improves model accuracy. Machine learning generally requires a large amount of data for training and validation, so collecting more clinical data in the future would likely enhance the accuracy and reliability of deep learning models.

In addition, interpretable AI techniques revealed that the heart and spine regions in X-rays are particularly significant, while certain lung areas show strong differences between infected and non-infected cases. It was also noted that the COVID cases in the dataset likely represent patients with severe symptoms, which explains the pronounced lung abnormalities. Consequently, the model may not perform as well for mild or asymptomatic cases.

In summary, the proposed framework demonstrates strong predictive capability and improves model trustworthiness. Interpretable AI can also assist healthcare professionals in clinical decision-making. However, practical application would require consideration of larger datasets and inclusion of both mild and severe infection cases.

9 Bibliography

- Ahsan, M. M., Gupta, K. D., Islam, M. M., Sen, S., Rahman, M., & Shakhawat Hossain, M. (2020). COVID-19 Symptoms Detection Based on NasNetMobile with Explainable AI Using Various Imaging Modalities. *Machine Learning and Knowledge Extraction, 2*(4), 490-504.
- Chang, M. (2016, October 27). Applied Deep Learning 11/03: Convolutional Neural Networks. Slideshare. https://www.slideshare.net/ckmarkohchang/applied-deep-learning-1103-convolutional-neural-networks
- Chollet, F. (2020, April 26). Grad-CAM: Class Activation Visualization. Keras. https://keras.io/examples/vision/grad_cam/
- IT People. (2020, January 19). Classification Algorithm: Multi-layer Perceptron. https://iter01.com/454359.html
- JT. (2018, April 16). VGG Deep Learning Principles. https://danjtchen.medium.com/
- Sun, H.-E. (2020, July 4). Understanding CNN Attention Regions: Introduction to CAM and Grad-CAM. AI Taiwan Academy. https://medium.com/ai-academy-taiwan/
- Tseng, K.-S. (2017, December 27). LIME - Local Interpretable Model-Agnostic Explanation: Technical Introduction. https://medium.com/@kstseng/
- Ning Wei, J. (2017, October 14). Image Processing: Superpixel Segmentation Using SLIC Algorithm. CSDN. https://blog.csdn.net/JNingWei/article/details/78236098