

Enhancing Chest X-Ray Image Classification for Lung Diseases through Machine Learning

Sungju Park¹, Hyunseo Cho¹, Ahyoung Chung¹, Timothy Han¹, Taeoh Yi¹, Angela Paik¹, and Taeheon Lee¹

¹Seoul Scholars International

ABSTRACT

Chest X-ray imaging is a widely used diagnostic tool for the detection and classification of various lung diseases. In this study, we propose a methodology to enhance the classification accuracy of chest X-ray images by leveraging machine learning techniques. Specifically, we employ the VGG-19 architecture to classify chest X-ray images into five distinct lung conditions: normal, tuberculosis, COVID-19, lung opacity, and pneumonia. A comprehensive dataset consisting of 31,787 chest X-ray images sourced from multiple medical institutions and hospitals worldwide is collected. Each image is labeled by expert radiologists with one of the five lung conditions. The dataset is then preprocessed and trained with 30 epochs. The trained VGGnet model achieved overall test accuracy of 95.11%, demonstrating its capability to accurately classify chest X-ray images into the five targeted lung conditions. The proposed methodology holds significant potential for improving the efficiency and accuracy of lung disease diagnosis based on chest X-ray images. By employing machine learning techniques, we can automate the classification process, providing clinicians with valuable decision support and expediting treatment plans. Moreover, the developed model can assist in the identification of lung diseases, including critical conditions such as COVID-19, enabling prompt and effective patient management.

Introduction

Chest radiographs, commonly known as X-rays, have been a fundamental diagnostic tool in the field of medicine, particularly in general practice. These images provide valuable insights into various indications and significantly impact patient management (Speets, 2006). The interpretation of chest X-rays has traditionally relied on the expertise of radiologists, which can be time-consuming and subjective. However, with the advent of computer-aided diagnosis (CAD) systems and advancements in machine learning, there has been a remarkable opportunity to enhance the efficiency and accuracy of chest X-ray image interpretation to classify lung diseases.

Machine learning techniques have shown great promise in the field of medical imaging, specifically in the analysis of chest X-ray images. By leveraging large-scale chest X-ray image datasets and training powerful artificial intelligence (AI) models, researchers can develop a robust Chest X-ray Image Classification System that holds the potential to revolutionize the diagnosis and treatment of lung diseases. Integrating machine learning techniques into CAD systems allows for automated identification and classification of lung diseases based on chest X-ray images, leading to improved accuracy and efficiency in diagnosis.

One of the key advantages of using machine learning in chest X-ray image classification is its ability to analyze large amounts of data and extract meaningful patterns and features. These AI models can learn from chest X-ray images to recognize specific visual characteristics associated with different lung pathologies such as pneumonia, tuberculosis, lung cancer, and chronic obstructive pulmonary disease (COPD). The development of accurate and reliable AI models in chest X-ray image classification can assist radiologists in their decision-making process, providing them with additional support and reducing the likelihood of misdiagnosis.

Several studies have demonstrated the effectiveness of machine learning approaches in chest X-ray image classification. For instance, Rajpurkar et al. developed a deep learning algorithm called CheXNet, which outperformed radiologists in detecting pneumonia from chest X-rays (2017). Similarly, using chest X-ray images, Lopes et al., employed machine learning techniques to differentiate between COVID-19 and other lung diseases accurately (2021). These studies highlight the potential of machine learning algorithms in improving the accuracy and efficiency of lung disease diagnosis based on chest X-ray images.

The development and implementation of a chest x-ray image classification system using machine learning techniques have the potential to transform the field of radiology and improve patient outcomes. By automating the identification and classification of lung diseases, clinicians can make more informed decisions and initiate appropriate treatment plans in a timely manner.

Therefore, the main research question of this study is how the integration of machine learning techniques into CAD systems can enhance the efficiency and accuracy of chest X-ray image interpretation for the classification of lung diseases. Based on this research question, our hypothesis is that if machine learning algorithms are integrated into CAD systems for chest X-ray image classification, then the accuracy and efficiency of lung disease diagnosis will significantly improve compared to traditional radiologist interpretation.

The prediction stemming from this hypothesis is that the automated identification and classification of lung diseases based on chest X-ray images, facilitated by machine learning algorithms within CAD systems, will yield higher accuracy rates and reduced diagnostic errors. It is expected that the use of machine learning techniques will improve the efficiency of diagnosis by reducing the time required for interpretation and providing radiologists with additional support in decision-making.

In this study, we aim to explore the integration of machine learning techniques into CAD systems for the automated identification and classification of lung diseases based on chest X-ray images. By investigating the potential improvements in efficiency and accuracy, we seek to contribute to the advancement of diagnostic capabilities and ultimately enhance patient outcomes in the field of radiology.

Research Design and Methods

Collection of Chest X-ray Images

To build an effective training dataset for this study, we gathered and merged five separate chest radiograph data sets from various medical institutions and hospitals globally. The dataset comprises a total of 31,787 chest X-ray images; each assigned a specific classification based on the diagnosis provided by radiologists. The categorized images are breakdown: 15,275 images depicting normal chest X-rays, 4,182 images showing Covid-19 cases, 5,618 images indicating pneumonia, 700 images displaying tuberculosis, and 6,012 images illustrating lung opacity. The datasets used in this study are detailed below.

COVID-19 Image Data Collection: Prospective Predictions Are the Future Data Set

This data set consists of 679 frontal chest X-ray images from 412 people from 26 countries that are diagnosed COVID-19. The dataset also contains metadata including clinical attributes about survival, ICU stay, intubation events, blood tests and more (Cohen, J.P., 2020). This study only used images of chest X-ray for image classification.

COVID-QU-Ex Dataset

The researchers of Qatar University have compiled the COVID-QU-Ex dataset, which consists of 3616 X-ray images of COVID-19, 1345 images of Viral Pneumonia, 6012 images of lung opacity and 10,192 images of normal X-rays. Lung segmentation masks are also provided for the entire dataset (Chowdhury, M. E., 2020). This study used images of COVID-19, pneumonia, lung opacity, and normal X-rays.

Tuberculosis (TB) Chest X-ray Database

Researchers from Qatar University, Doha, Qatar, and the University of Dhaka, Bangladesh, along with their collaborators from Malaysia and medical doctors from Hamad Medical Corporation and Bangladesh, have created a database of chest X-ray images for Tuberculosis (TB) positive cases and Normal images. There are 700 tuberculosis images and 3500 normal images(Rahman, T., 2020).

Labeled OCT and Chest X-ray images dataset

There are both OCT and chest X-ray images provided in the dataset. Specifically, the chest X-ray images (anterior-posterior views) were selected from retrospective cohorts of pediatric patients aged one to five years old at Guangzhou Women and Children's Medical Center in Guangzhou. The dataset consists of 5,863 X-ray images, classified as either pneumonia or normal cases. All chest X-ray imaging was conducted during the patient's routine clinical care. To ensure data quality, a rigorous screening process was implemented to remove low-quality or unreadable scans. Two expert physicians independently graded the diagnoses for the chest X-ray images, ensuring accuracy before training the AI system. Additionally, a third expert reviewed the evaluation set to account for any potential grading errors(Kermany, 2018). This study only utilized chest x-ray images.

Actualmed COVID chest x-ray dataset

A total of 92 chest X-ray images, consisting of 12 COVID-19 images and 80 normal cases, were compiled by researchers from DarwinAI Corporation, Canada, and the Vision and Image Processing Research Group at the University of Waterloo, Canada. In addition to the images, accompanying doctor's notes were also provided for each image as a metadata(Chung, A., 2020).

Data Preprocessing

Image Classification

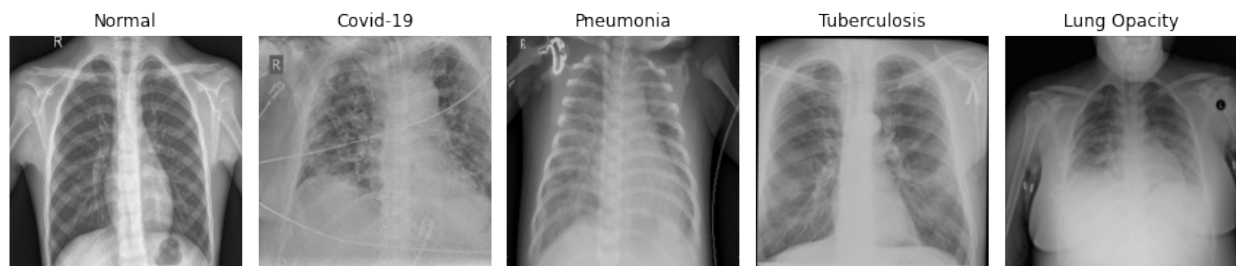


Fig 1. Chest X-ray Images Multi Classification

A dataset consisting of 31,787 chest X-ray images was classified into five distinct categories: normal, COVID-19, pneumonia, tuberculosis, and lung opacity. Images that did not fall into these predefined categories were excluded from the dataset. To facilitate further processing, the category labels were converted into numerical values using label encoding. Specifically, normal photographs were assigned the value 0, COVID-19 was assigned 1, pneumonia was assigned 2, tuberculosis was assigned 3, and lung opacity was assigned 4. Each class exhibited different visual features.

Image Processing and Normalization

Given that the chest X-ray images were collected from various data sets and had different sizes, they were resized to a uniform dimension of 256 pixels in height and 256 pixels in width, ensuring that important information was not lost during the process. To integrate the datasets, both colorized and grayscale images were combined. To ensure

consistency during training, the RGB values were adjusted, and all images were converted into grayscale. The RGB values, ranging from 0 to 255, were divided by 255 to normalize them within the range of 0 to 1, making them suitable for further processing in the model. Additionally, image augmentation techniques such as 180° rotation and filling adjacent empty spaces with reversed images were applied to increase the number of images and enhance the consistency of the dataset.

Model Structure

```
Model: "model_4"
```

Layer (type)	Output Shape	Param #
input_13 (InputLayer)	[(None, 256, 256, 3)]	0
vgg19 (Functional)	(None, 8, 8, 512)	20024384
global_average_pooling2d_4 ((None, 512)		0
dense_16 (Dense)	(None, 5)	2565

```
Total params: 20,026,949
Trainable params: 20,026,949
Non-trainable params: 0
```

Fig 2. Composition of Model Layers

For the classification task, the study utilized the VGG-19 network as the model architecture. The VGGNet is a convolutional neural network (CNN) model that was originally developed for image recognition tasks (Simonyan & Zisserman, 2014). It has achieved remarkable performance in various computer vision tasks due to its deep architecture and use of small convolutional filters.

The VGG-19 network consists of 19 layers, including 16 convolutional layers, 3 fully connected layers, and max pooling layers in between (Simonyan & Zisserman, 2014). Each convolutional layer is followed by a rectified linear unit (ReLU) activation function, which introduces non-linearity into the model and allows for better feature representation.

By applying the VGG-19 network to the chest X-ray images, the model learns hierarchical features at different levels of abstraction. The deep layers of the VGG-19 network capture complex patterns and spatial information, enabling the model to distinguish between different lung conditions based on the visual characteristics present in the chest X-ray images.

Experimental Results

Implementation Details

The training was carried out by MacBook Pro (13-inch, 2019, Two Thunderbolt 3 ports) PC. 1.4 GHz Quad-Core Intel Core i5 was used for the CPU, and Intel Iris Plus Graphics 645 1536 MB was used for GPU. Jupyter Notebook was executed via the anaconda python compiler. In realizing the model, Python, and Pandas were used for data storage. Python-OpenCV was used for image conversion. TensorFlow and Keras were used to create model layers, while Scikit Learn was used to assort data. 31,787 chest x-ray images were created in a data frame and saved (used Pandas), and to verify the deep learning model, the Scikit Learn library was used to train, test, and validate data at the ratio of 7:2:1. The assorted images were adjusted to batch size 32, 256* height, and 256* width, using Python OpenCV library.

Accuracy and Loss Rate

Accuracy refers to the rates of correct predictions out of the entire data set and is the most common criterion for assessing the model. The overall accuracy of the model was 95.11% and the accuracy of learning and verification is set out in Figure 2 below.

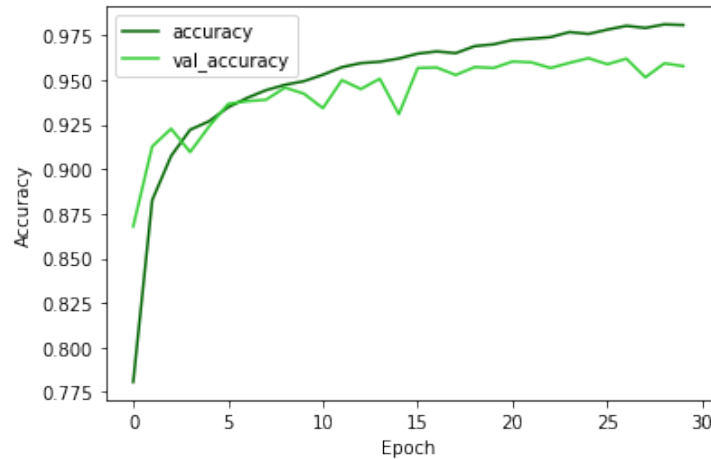


Fig 3. Accuracy of Training & Validation dataset

The loss rate refers to the margin of error between the predicted and actual values in the model. The bigger the loss is, the larger the margin becomes. The loss rate was 21.23% in this model.

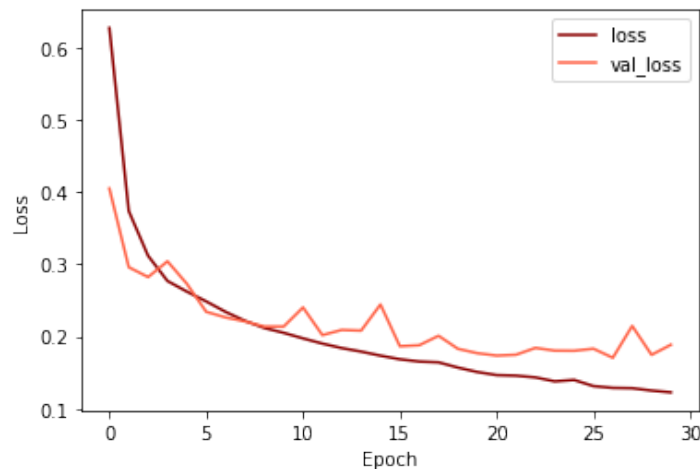


Fig 4. Loss rate of Training & Validation dataset

Confusion Matrix

The Confusion Matrix provides insights into the classification performance of the VGG19 model. A detailed analysis found that the rate of correctly classifying normal X-ray images as normal was 98.79%. For COVID-19 cases, the accuracy in correctly classifying them was 95.11%. The classification accuracy for pneumonia cases was 96.44%.

Tuberculosis cases were classified accurately at a rate of 95.71%. However, when it came to identifying lung opacity in X-ray images, the accuracy dropped to 83.96% in correctly classifying such cases.

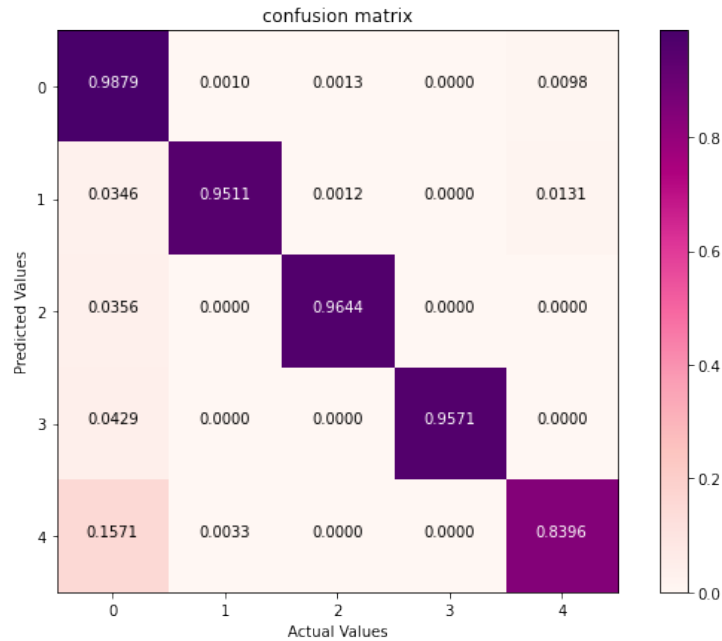


Fig 5. Confusion Matrix

ROC and AUC

The Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) measure the performance of a classification model. In the analysis, the ROC AUC value was found to be 96.15%. While the other classes achieved high AUC rates, the class for Lung Opacity (label 4) had a slightly lower rate in comparison.

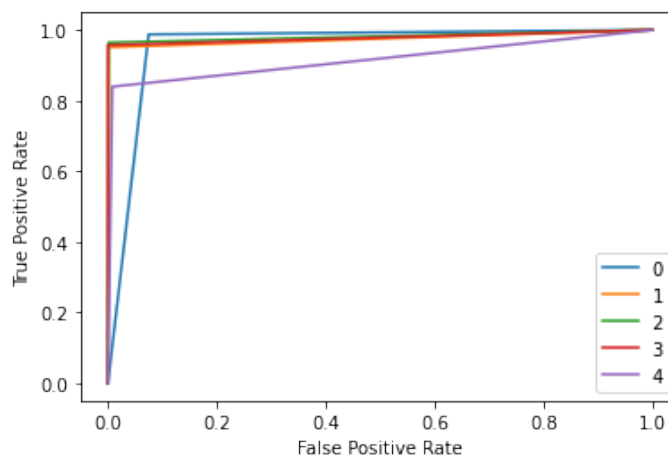


Fig 6. ROC curve

Classification Report

Based on the classification report, the precision rate for each classification was consistently above 92%. Regarding the recall rate, all classes except for label 4 (Lung Opacity) achieved over 95%. Label 4 had a recall rate of 84%. Furthermore, when considering the f1 score, all classes obtained scores exceeding 90%.

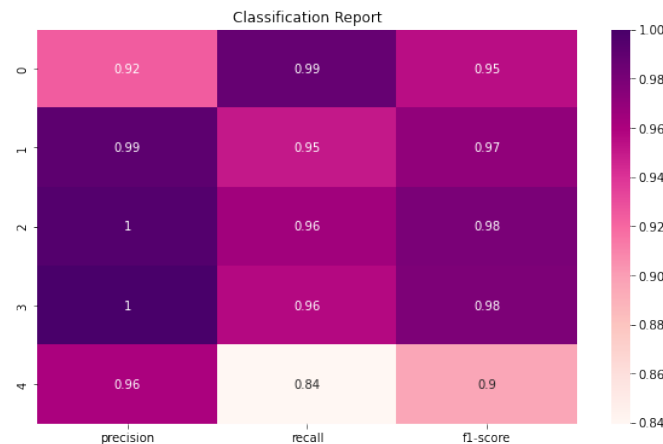


Fig 7. Classification report

Sensitivity and Specificity

The sensitivity rate, also known as the true positive rate, was found to be 95.11%. This indicates the model's ability to correctly identify positive instances. The specificity rate, also known as the true negative rate, was 98.79%, representing the model's ability to accurately identify negative instances.

Conclusion and Justification

In conclusion, this study focused on enhancing the classification of chest X-ray images for lung disease diagnosis through machine learning techniques. Using the VGGnet model, the paper aimed to accurately classify chest X-ray images into five different lung conditions: normal, tuberculosis, COVID-19, lung opacity, and pneumonia. The study addressed the limitations of traditional radiologist interpretation and aimed to improve the efficiency and accuracy of lung disease diagnosis using automated machine learning algorithms.

The significance of this study lies in its potential to revolutionize the field of radiology and improve patient outcomes. By integrating machine learning techniques into computer-aided diagnosis (CAD) systems, the accuracy and efficiency of chest X-ray image interpretation can be significantly enhanced. This can lead to improved diagnostic accuracy, timely initiation of treatment plans, and ultimately better patient management.

The justification for further research and development in this area is based on the promising results demonstrated by previous studies using machine learning in chest X-ray image classification. Research has shown that machine learning algorithms, such as the VGGnet model, can outperform radiologists in detecting lung diseases from chest X-rays. By building upon these findings and continuing to refine the models, there is a great potential to improve the accuracy and efficiency of lung disease diagnosis.

Funding for this research is warranted due to its potential to bring changes in the field of medical imaging and radiology. The integration of machine learning techniques into CAD systems has the potential to enhance the diagnostic capabilities of healthcare professionals, reduce the likelihood of misdiagnosis, and improve patient

outcomes. Investing in further research, data collection, and model refinement will contribute to the advancement of chest X-ray image classification and benefit both healthcare providers and patients.

In summary, the development and application of machine learning techniques in the classification of chest X-ray images for the diagnosis of lung diseases offer significant potential for improving healthcare practices and patient care. By addressing the limitations of traditional interpretation methods and leveraging the power of machine learning algorithms, this research has the capacity to transform the field of radiology and enhance the accuracy and efficiency of lung disease diagnosis. Continued research and funding in this area are essential to harness the full potential of machine learning in chest X-ray image classification and its impact on healthcare outcomes.

References

- Chowdhury, M. E., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M. A., Mahbub, Z. B., Islam, K.R., Khan, M. S., Iqbal, A., Emadi, N. A., Reaz, M. B., & Islam, M. T. (2020). Can ai help in screening viral and covid-19 pneumonia? *IEEE Access*, 8, 132665–132676. <https://doi.org/10.1109/access.2020.3010287>
- Chung, A. Actualmed COVID-19 chest x-ray data initiative. <https://github.com/agchung/Actualmed-COVID-chestxray-dataset> (2020)
- Cohen, J. P., Morrison, P., Dao, L., Roth, K., Duong, T., & Ghassem, M. (2020a). Covid-19 image data collection: Prospective predictions are the future. *Machine Learning for Biomedical Imaging*, 1(December 2020), 1–38. <https://doi.org/10.59275/j.melba.2020-48g7>
- Kermany, Daniel; Zhang, Kang; Goldbaum, Michael (2018), “Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification”, *Mendeley Data*, V2, doi: 10.17632/rsbjbr9sj.2
- Lopes, C., Martins, J., Cavadas, V., Lourenço, A., & Campilho, A. (2021). COVID-19 classification from chest X-ray images using deep learning and transfer learning. *Diagnostics*, 11(4), 700.
- Rahman, T., Khandakar, A., Kadir, M. A., Islam, K. R., Islam, K. F., Mazhar, R., Hamid, T., Islam, M.T., Kashem, S., Mahbub, Z. B., Ayari, M. A., & Chowdhury, M. E. (2020). Reliable tuberculosis detection using chest X-ray with deep learning, segmentation and visualization. *IEEE Access*, 8, 191586–191601. <https://doi.org/10.1109/access.2020.3031384>
- Rahman, T., Khandakar, A., Qiblawey, Y., Tahir, A., Kiranyaz, S., Abul Kashem, S. B., Islam, M. T., Al Maadeed, S., Zughailer, S. M., Khan, M. S., & Chowdhury, M. E. H. (2021). Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images. *Computers in Biology and Medicine*, 132, 104319. <https://doi.org/10.1016/j.combiomed.2021.104319>
- Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017, December 25). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. *arXiv.org*. <https://arxiv.org/abs/1711.05225>
- Simonyan, K., & Zisserman, A. (2015, April 10). Very deep convolutional networks for large-scale image recognition. *arXiv.org*. <https://arxiv.org/abs/1409.1556>

Speets, A. M. (2006). Upper abdominal ultrasound in general practice: Indications, diagnostic yield and consequences for patient management. *Family Practice*, 23(5), 507–511.
<https://doi.org/10.1093/fampra/cml027>