Pneumonia Detection in Chest X-Rays Using Deep Learning

Vedant Valluri

Veritas AI, Wilmington, Delaware, USA, 19801

BRIEF. This study applies artificial intelligence to detect pneumonia in chest X-rays.

ABSTRACT. In this work, we aim to detect pneumonia in chest X-rays using Deep Learning (DL) based Artificial Intelligence models. The data used for this study are Chest X-ray images from the RSNA Pneumonia Detection challenge. Each image is labeled with one of three class labels, (No Pneumonia (0), Pneumonia (1), and No Pneumonia / Not Normal (2)). One of the challenges considered here is the inclusion of Class 2 (Not Pneumonia / Not Normal), indicating patients who didn't have pneumonia but had some other affliction. In this study, we trained models for two classification tasks: a binary classification task (No Pneumonia versus Pneumonia) and a 3-class classification task (No Pneumonia, Pneumonia, and No Pneumonia / Not Normal). Given the effectiveness of Convolutional Neural Networks (CNNs) in analyzing image data, we experiment with multiple CNN architectures for this task. Amongst the different CNN architectures we tried, the ResNet50 model achieved the highest accuracy, achieving 73.4% accuracy for the 3-class classification and 94% accuracy for binary classification. We hypothesize that the diversity of afflictions presented in images of Class 2 made the 3-class classification more challenging. Given the high accuracy of the binary classification model, we conclude that DL-based models can be a useful tool to help detect pneumonia in chest X-rays.

INTRODUCTION.

Pneumonia Context.

Pneumonia is a significant global health concern, responsible for over 15% of deaths in children under 5 years old worldwide, with 920,000 fatalities reported in 2015. In the United States, pneumonia results in more than 500,000 emergency department visits [1] and over 50,000 deaths in 2015 [2], ranking it among the top 10 causes of death in the country. Accurately diagnosing pneumonia is challenging as it requires the expertise of specialists who review chest radiographs (CXRs) along with clinical history, vital signs, and laboratory exams for confirmation [3]. CXRs are commonly used for diagnosis of pneumonia, but their interpretation is complicated by the presence of other lung conditions that exhibit similar opacities, such as pulmonary edema, atelectasis, lung cancer, and pleural effusion. Factors like patient positioning and inspiration depth further complicate CXR interpretation [4], and the high volume of images clinicians need to read during their shifts adds to the complexity of pneumonia diagnosis.

Computer Vision and CNNs.

Computer Vision (CV) involves the utilization of Artificial Intelligence (AI) techniques to interpret and extract meaningful information from visual data, such as images and videos [5]. CNNs, a fundamental component of CV, are DL models structured with convolutional and pooling layers for automatic hierarchical feature extraction [6]. CNNs' key innovation lies in parameter sharing and convolution, enabling them to recognize patterns and features efficiently throughout an image [6]. These networks are trained via backpropagation, where they adjust internal parameters to minimize prediction errors, and have excelled in numerous CV tasks, including image classification, object detection, and facial recognition [7][8]. CNNs have revolutionized CV, offering powerful tools to analyze and understand visual data, mirroring the hierarchical feature learning in the human visual system [6]. Given CNNs' success for CV tasks, they are a promising tool for automated classification/interpretation of Chest X-rays.

Past Work in Computer Vision for Pneumonia.

There has been a lot of past work in attempting to use CNNs / CV for the task of pneumonia detection. Some work introduces a CNN-based model employing Dynamic Histogram Equalization (DHE) to enhance image contrast, achieving a remarkable accuracy of 96.07% and a precision rate of 94.41%, surpassing existing CNN models [9]. Other work has used more complicated models such as DenseNet201, outperforming other similar CNN models [10]. One review paper discusses advancements in AI-based approaches for COVID-19 detection from chest X-ray images, emphasizing high accuracy using models like VGG-19 and ResNet, while also highlighting the need for expanded databases and improved classification techniques [11]. These studies collectively underscore the substantial progress and potential applications of AI in pneumonia diagnosis. Building upon this previous work, we experiment with CNN-based models for 3-class and binary classification tasks.

Given the above context, the goal of this study is to apply computer vision and deep learning methods to the problem of pneumonia detection. Specifically, given the focus on binary classification in past work, we sought to explore 3-class classification, and compare the performance between this and binary classification. The reason to study the 3-class models is because some patients might present with other afflictions which aren't pneumonia but are still not normal.

The next section provides an overview of the data used in this study, the various CNN models explored and an overview of the training approach. The following section outlines the results achieved for the 3-class and binary classifications tasks. In the final section we discuss the interpretation of the results and potential future direction this project can take.

METHODS.

Data.

The dataset used for this project was from the Radiological Society of North America (RSNA) Pneumonia Detection Challenge [12]. RSNA worked in collaboration with the Society for Thoracic Radiology and MD.ai to create labels for chest X-rays made public by the National Institutes of Health (NIH) [12]. It contained around 26,000 samples of Chest X-rays as DICOM images along with labels for three classes (No Pneumonia, Pneumonia, and No Pneumonia/Not Normal) and bounding box values for those samples containing pneumonia. DICOM (Digital Imaging and Communications in Medicine) is a standard for storing, transmitting, and sharing medical images and related information, such as X-rays and MRI scans, in a standardized digital format [13]. It ensures compatibility and interoperability among different medical imaging devices and software systems. The images were 1024x1024 pixels and were sized down to 224x224 due to the model's input size requirements. The provided data had already been split as train and test, but it was decided to further split the train dataset into an 80/20 partition randomly for train and validation respectively. We trained the model on the train dataset and report evaluation metrics on the validation set. The labeled test set was held for the Kaggle competition, and so we were unable to run our final model on the test set.

Table 1. Results from the base models for 3-Class Model

	Accuracy	AUC	Precision	F1 Score
ResNet50	73.75%	0.9003	74.34%	72.55%
VGG-16	71.48%	0.8796	72.70%	70.20%
VGG-19	72.46%	0.8879	73.33%	71.40%

Models and Training.

One of the questions to be considered when working with a dataset such as this is whether to use the third class (No Pneumonia/Not Normal) as a separate class or to combine it with the first one (No Pneumonia), since both represent images without pneumonia. We settled on keeping them separate and studying the three-class problem first. We tried multiple models (ResNet50, VGG-16, VGG-19) as base models for the CNN whose performances are shown below.

It can be seen that out of these models the ResNet50 is the best performing with the VGG-19 only slightly worse. Due to these findings the rest of the study is based on ResNet50.

ResNet (Residual Network) [14] is a deep convolutional neural network architecture introduced by Microsoft Research in 2015. It is known for its innovative use of residual blocks, which allow for very deep networks to be trained more effectively by mitigating the vanishing gradient problem. ResNet has had a significant impact on CV tasks and is widely used for image classification and object detection. VGG (Visual Geometry Group) [15] is another deep convolutional neural network architecture developed by the University of Oxford in 2014. VGG is characterized by its simplicity and uniform architecture, consisting of 16 or 19 weight layers with small 3x3 convolutional filters and max-pooling layers. Despite its simplicity, VGG achieved competitive performance on various image recognition tasks and served as a foundational model for DL research [11].

Our models were run using a batch size of 64 and a learning rate of 0.0005. The categorical cross-entropy loss function was used along with the AdamW optimizer, and the training was run for 30 epochs. The training was done on a V100 GPU running on Google Collaboratory. The accuracy plateaued after around 20 epochs. The first approach tried was to train the model for double the number of epochs, to see if more training could improve the accuracy. This proved fruitless as more training improved the accuracy of the model by less than a percent and in some cases made the performance worse, likely due to overfitting.

We studied models for two separate classification tasks: (A) a 3-class classification task which included the entire training dataset with three classes (No Pneumonia (0), Pneumonia (1), and No Pneumonia / Not Normal (2)); (B) a binary classification task which only included training data for two classes (No Pneumonia (0), Pneumonia (1)).

RESULTS.

Three Class Classification.

We achieved an accuracy of 73.75% on the 3-class classification task using ResNet50 as shown in Table 1. Given the low accuracy we wanted to study what the model was predicting for each of the samples. To study this we created a matrix to assess what percent of each of the real classes was going into each of the predicted classes. Table 2 below shows this matrix for the 3-class ResNet50. The rows represent the actual label, and the columns are predicted values.

The matrix shows that 88.03% of the true Class 0s (No Pneumonia) were predicted as a 0 and 11.85% were predicted as Class 2 (No Pneumonia / Not Normal). For the real Class 1s (Pneumonia), 55.74% were predicted correctly as 1s and 41.15% were incorrectly predicted as Class 2. Almost 30% of the real Class 2s were incorrectly classified as either 0 (15.06%) or 1 (13.37%), and 71.57% were correctly predicted as Class 2. From the results, we gather that one of the main issues with this model is that it was unable to accurately distinguish between Class

Table 2. Confusion Matrix for 3-Class Model

	Predicted No Pneumonia	Predicted Pneumonia	Predicted No Pneumonia/Not Normal
True No Pneumonia	88.03%	0.11%	11.85%
True Pneumonia	3.10%	55.74%	41.15%
True No Pneumonia /Not Normal	15.06%	13.37%	71.57%

1 (Pneumonia) and Class 2 (No Pneumonia / Not Normal). This led to the idea of creating a second stage model that would take the predicted 1s and 2s from this model and run them through a second model that was trained on these 2 classes (Class 1, Class 2) and excluded Class 0 data. This model had bad results. While it was correctly predicting most of the Class 2s into their correct class, it was incorrectly predicting more than half of the Class 1s as Class 2. From this, we hypothesized that the model was unable to learn how to accurately handle Class 2 because the images classified under Class 2 could have represented a wide range of conditions, some of which might have had similar image characteristics as pneumonia.

Binary Classification.

Given the complications of handling Class 2 (No Pneumonia / Not Normal) we wanted to see how the model would perform when it was only trying to distinguish between images that had No Pneumonia (Class 0) and images that had Pneumonia (Class 1). This was a binary classification problem where Class 2 data was excluded, and the model was trained only on Class 0 and Class 1 data.

This model performed extremely well compared to the 3-class task and had an overall accuracy of 94%, an AUC of 0.9855, precision of 93.73% and an F1 Score of 93.54%.

To further test the model to represent a real-world setting more accurately, we presented it with Class 2 images that it hadn't been trained on. For these images, the model predicted 'Pneumonia' 73% of the time and 'No Pneumonia' 27% of the time. From this we hypothesized that from the model's perspective, many of the Class 2 images could have had a condition where the CXR image had similar characteristics to CXRs with pneumonia.

In summary, the study aimed to detect pneumonia in chest X-rays using deep learning techniques on a dataset with three classes: No Pneumonia (0), Pneumonia (1), and No Pneumonia/Not Normal (2). ResNet50 emerged as the most effective model, achieving 73.75% accuracy for the 3-class classification and 94% accuracy for the binary classification (No Pneumonia vs. Pneumonia).

DISCUSSION.

These results show that when distinguishing between No Pneumonia (0) and Pneumonia (1), a DL-based model achieves a high accuracy of 94%. However, the inclusion of Class 2 (No Pneumonia / Not Normal) "confuses" the model and reduced the accuracy to 73.75%. This indicates an inherent difficulty in differentiating Class 2 from the other two classes. In the 3-class problem, analysis highlights that the more daunting challenge was distinguishing between Class 1 (Pneumonia) and Class 2 (No Pneumonia / Not Normal), indicating that images of Class 1 and Class 2 probably had similar "image characteristics". This made it harder for the model to learn how to accurately distinguish between these two classes. The results also show that data from Class 0 (No Pneumonia) and Class 1 (Pneumonia) exhibit distinct characteristics, making them easier to classify. We conclude that a binary classification model trained using an appropriate dataset consisting of 'No Pneumonia' and 'Pneumonia' images can serve as a useful tool in helping to detect pneumonia in Chest X-rays. However, in practice,

such models alone will be insufficient. As we discussed earlier, such a model could have prediction errors when presented with images that might have similar characteristics to images with pneumonia but aren't pneumonia.

To address this challenge and further improve classification accuracy, future work should delve deeper into data analysis, identifying key features that distinguish between classes 0, 1, and 2. Also, if a second stage model is built which can place bounding boxes around the regions in the image suspected to be pneumonia, then it could help to more accurately distinguish between Class 1 and Class 2 images. Class 1 images would have one or more bounding boxes while Class 2 images might not have any bounding boxes due to the lack of pneumonia.

Additionally, data augmentation techniques could be employed to diversify the samples within Class 2, and ensemble methods might be explored to leverage the strengths of different models for different classes. This research underscores the importance of tailored strategies for handling distinct classes within a dataset to enhance the performance of machine learning models in complex classification tasks.

REFERENCES.

1. P. Rui, K. Kang, National Ambulatory Medical Care Survey: 2015 Emergency Department Summary Tables. Table 27. Available from: www.cdc.gov/nchs/data/nhamcs/web_tables/2015_ed_web_tables.pdf, Jul 2023.
2. Deaths: Final Data for 2015. Supplemental Tables. Tables I-21, I-22. Available from: www.cdc.gov/nchs/data/nvsr/nvsr66/nvsr66_06_tables.pdf, Jul 2023.
3. T. Franquet, Imaging of community-acquired pneumonia. *J Thorac Imaging* **2018**, 282-294 (2018).
4. B. Kelly, The Chest Radiograph. *Ulster Med Journal* **81(3)**, 143-148 (2012).
5. Y. LeCun, Y. Bengio, & G. Hinton, Deep learning. *Nature* **521**, 436-444 (2015).
6. A. Krizhevsky, I. Sutskever, & G. E. Hinton, ImageNet classification with deep convolutional neural networks. *Communications of the ACM* **60(6)**, 84-90 (2017).
7. I. Goodfellow, Y. Bengio, A. Courville, & Y. Bengio, Deep learning (Vol. 1). MIT press Cambridge (2016).
8. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions. *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 1-9 (2015).
9. D. Zhang, F. Ren, Y. Li, L. Na, Y. Ma, Pneumonia Detection from Chest X-Ray Images Based on Convolutional Neural Network. *Electronics*, **10**, 13, 1512 (2021).
10. T. Rahman, M. E. H. Chowdhury, A. Khandakar, K. R. Islam, K. F. Islam, Z. B. Mahbub, M. A. Kadir, S. Kashem. "Transfer Learning with Deep Convolutional Neural Network (CNN) for Pneumonia Detection Using Chest X-ray." *Applied Sciences* **10**, 9 (2020).
11. I. Muhammad, H. Rehman, A. Nait-ali. 'Detection of Covid-19 From Chest X-Ray Images Using Artificial Intelligence: An Early Review'. arXiv:2004.05436 (2020).
12. "RSNA Pneumonia Detection Challenge (2018)." RSNA, RSNA, www.rsna.org/education/ai-resources-and-training/ai-image-challenge/rsna-pneumonia-detection-challenge-2018.
13. M. Mustra, K. Delac, M. Grgic, Overview of the DICOM standard. *IEEE* **2008**. 39 – 44 (2008).
14. K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 770-778 (2016).
15. K. Simonyan, A. Zisserman. "Very deep convolutional networks for large-scale image recognition." *International Conference on Learning Representations* (2015).

Vedant Valluri is a student at Monta Vista High School in Cupertino, CA; he participated in a research internship through the Veritas AI program.