A Convolutional Neural Network to Diagnose Pneumonia from Frontal Chest X-Ray Images

Kevin Mao
Columbia University, Computer Science
New York City, NY
kevin.mao@columbia.edu

Alex Kim
Columbia University, Biomedical Engineering
New York City, NY
agk2144@columbia.edu

***Abstract*—Pneumonia is the leading cause of death among children under 5 years old [1], and early diagnosis in critical for proper treatment. In this paper, we show that transfer learning applied to large convolutional neural networks can be used to diagnose pneumonia using a Kaggle dataset of 5,858 frontal chest X-ray images [8] from 5,232 unique children. We were able to replicate previous work and build on this by developing a multi-class model that could differentiate viral and bacterial pneumonia**

We developed two models, for differentiating between (1) normal and infected images, and (2) normal, bacterial, and viral pneumonia images. Our binary algorithm achieved a best F1 test score of 0.941 and accuracy 91.441% with three hidden layers, batch size of 64, learning rate of 0.001, and 50 epochs. This performance was within a similar range as previous work with this dataset [7]. Our multi-classifying algorithm achieved a best F1 test score of 0.636 and accuracy 83.784% with three hidden layers, batch size of 1000, learning rate of 0.001, and 400 epochs.

We have shown that our model is capable of binary and 3-class classification of chest X-ray images for pneumonia, but more can be done to further improve performance.

***Index Terms*—Convolutional Neural Network, Pneumonia, Chest X-Ray, Biosignals, Deep Learning**

I. Introduction

According to a study done by UNICEF, pneumonia remains the leading cause of death among children under 5 years old [1]. In 2015, pneumonia was responsible for 922,000 child deaths, making up almost 16% of all child deaths [2]. More than 1 million people in the US alone are hospitalized from pneumonia every year [2]. Despite there being effective methods to diagnose and treat pneumonia, high poverty areas with suboptimal health conditions (malnutrition, contaminated water and air pollution) are greatly affected by the disease [1].

Pneumonia arises when inflammation occurs in the alveoli of lungs due to the presence of bacteria or viruses. These contaminants can be introduced via water droplets or food particles that enter the respiratory tract [3]. While healthy adults can recover fairly quickly, for younger and older patients with weakened immune systems pneumonia may be life threatening.

Methods of diagnosing pneumonia may range depending on the resources available. Diagnoses in low-income countries primarily rely on identifying symptoms early with simple clinical procedures, such as observing the breathing quality and rate of the patient using a stethoscope. Patients with access to more precise methods may have an x-ray of their lungs taken to identify which part of the lung tissue is inflamed [3].

There are several issues that remain with the status quo of radiology exams. A combination of low resources and high volume may lead to misclassifications and consequent misdiagnoses by WHO protocol trained radiologists. In fact, studies have shown there are discrepancies in the assessment of radiographs between radiologists (with concordance rates between two reviewers to be as low as 48%) [4]. Secondly, proper treatment of pneumonia depends on the correct identification of the pathogen. Antibiotics, like amoxicillin, are administered as a blanket drug for children pneumonia diagnoses in low-income regions, which may not be effective against viral pneumonia or other forms of bacteria with different site specificities.

Artificial intelligence has shown promise in surpassing the performance of medical professionals for classification tasks. Traditional methods of image analysis involved a triad approach of manually segmenting objects of interest, developing classifiers for each class of objects, and classifying the image [6]. Now, convolutional neural networks have become a standard architecture for processing and classifying images. Deep neural networks have shown an ability to quantify features that are indicative of the diseases like diabetes mellitus (from retinopathy images) [9] and skin cancer (from images of skin) with high accuracy [5].

For this study, we adopted a transfer learning approach which involves reusing a base network trained on a base dataset and repurposing the learned features to a dataset of interest. This process is made possible as deep NNs that are trained on images tend to learn similar features. The weights of this model learned on extremely large datasets like the Stanford ImageNet dataset can be leveraged and retrained on the last layer with back-propagation to recognize shared features in biomedical images like chest x-rays. In this project, we explore a transfer learning framework using Googles ImageNet model to effectively classify normal, viral, and bacterial pneumonia in pediatric chest X-rays images.

II. Materials and Methods

A. Problem Formulation

For the purposes of this experiment, two problems were considered. Given the nature of the dataset which was divided into normal and abnormal (bacterial and viral separately labelled) pneumonia images we decided to assess the performance of the network as both a binary (normal vs. abnormal) and multi label classifying (normal vs. bacterial vs. viral) problem. Input X is the provided chest X-ray images and the output Y is a label $y \in \{0, 1\}$ for the binary classifier and $y \in \{0, 1, 2\}$ for the multi-classifier. A binary cross entropy loss function was optimized for the binary classifier, and a sparse softmax cross entropy (categorical cross entropy loss) for the multi-classifer.

In order to combat potentially dangerous misdiagnoses, precision and recall performance metrics can be used to give false negatives higher costs than false positives. The F1-score accounts for both precision and recall scores, and is used to gauge the performance of our network. Accuracy is also recorded.

B. Model Architecture and Training

We used large convolutional neural networks that were pretrained on the ImageNet dataset and switched out their last classification layers for ours. We tried both Xception and InceptionV3 models, models that had been used in previous studies [9] [7], but found better results with Inception. After the pretrained model, we added a global average pooling layer, followed by a variable number of dense layers. The output after the global average pooling layer was saved and loaded in a separate model comprised only of dense layers, and this was the model we trained. This technique significantly sped up training times without sacrificing performance. We used an Adam Optimizer for training and found the best number of epochs, batch size, and training step empirically. Training was done on Google Cloud Deep Learning VM with a NVIDIA Tesla V100 GPU and all code was written in Tensorflow. Figures 4 and 5 tabulate the results of all the hyperparameters we tested on both classifiers.

C. Data

We use the Kaggle dataset Kermany et. al (2018) [7] which contains 5,585 frontal chest X-ray images of 5,232 unique children. Each image is annotated with pathology labels of normal, bacterial, and viral pneumonia. Refer to Figure 1 (taken from [7] for sample X-rays). There was no patient overlap between training, validation, and testing datasets.

Label analysis showed that the number of images in each folder were inherently unbalanced. There were 0 viral pneumonia images and only 17 images total in the validation data set, while the testing data set had 234, 242, and 148 images and the training data had 1342, 2530, 1345 images for normal, bacterial, and viral cases, respectively. In order to normalize the number of images, two measures were taken. First, validation and test data were combined, randomized, and split evenly between validation and test data sets. Secondly, a random selection of the minimum number of images for each label type was selected. Table 1, below shows the data before and after balancing the datasets.

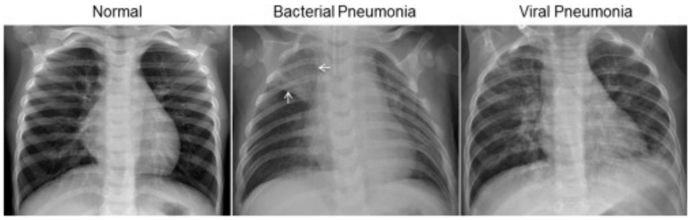


Fig. 1. Sample X-rays for each class [7]

All images from training, validation, and testing data were resized to 300x300 and converted into RGB to address shape discrepancies between images and to accomodate the input shape of InceptionV3 and Xception. We also augmented the training data by generate new samples by flipping images horizontally, brightening, and rotating by 20 degrees.

	Before Balancing			**After Balancing**		
	Normal	***Bacterial***	***Viral***	***Normal***	***Bacterial***	***Viral***
Train	1342	2530	1345	1342	1342	1342
Validate	9	8	0	74	74	74
Test	234	242	148	74	74	74

Table 1. Number of images for train, validate, and test sets per label, before and after pre-processing.

III. Results

We found that our best model for 3-class classification had 3 additional dense layers, a batch size of 1000 images, a learning rate of 0.001, and it was trained on 400 epochs. This model had a f1 score of 0.636 and an accuracy of 83.78% for 3 class classification. Additionally, the model predicted labels with above a 0.8 true positive rate. Refer to Figure 2 for the normalized confusion matrix and Figure 3 for the train and validation loss curves. These parameters were chosen to match previous work [7], but there were better models in binary classification.

A model with 3 dense layers, a batch size of 64, a learning rate of 0.001, and 50 epochs, reached a f1 score of 0.941 and an accuracy of 91.44% for binary classification.

IV. Discussion

As expected, our binary classifier performed better than our 3-classifier model because differences in features between viral and bacterial pneumonia images were very subtle. Our binary model performed similarly to Kermany et. al (2018) who used the same dataset and achieved 92.8% with a sensitivity of 93.2% and a specificity of 90.1% for their pneumonia vs. normal model [7]. For a bacterial vs. viral pneumonia model,

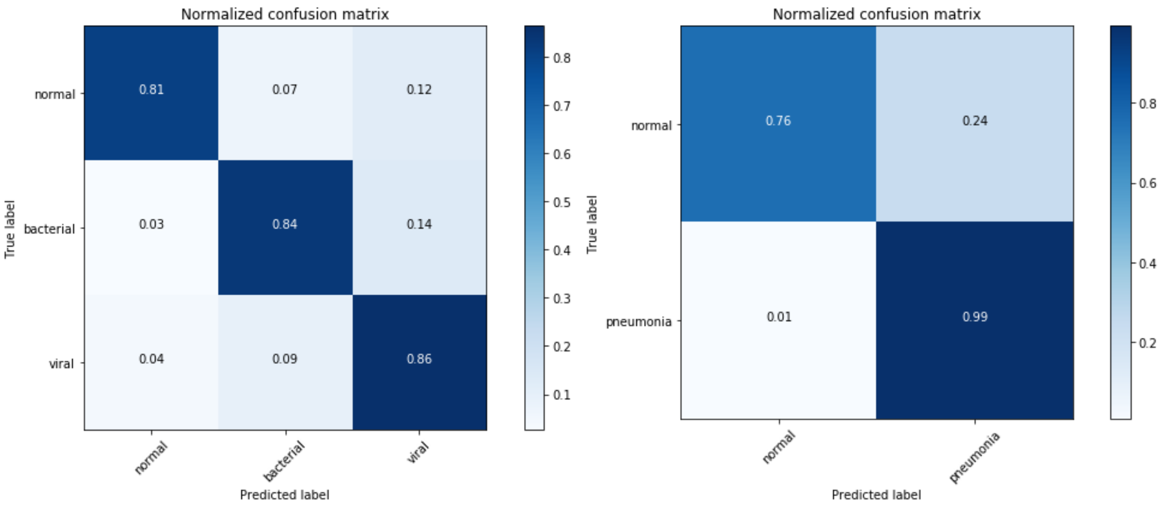


Fig. 2. Normalized confusion matrix for (left) multi-class model and (right) binary-class model.

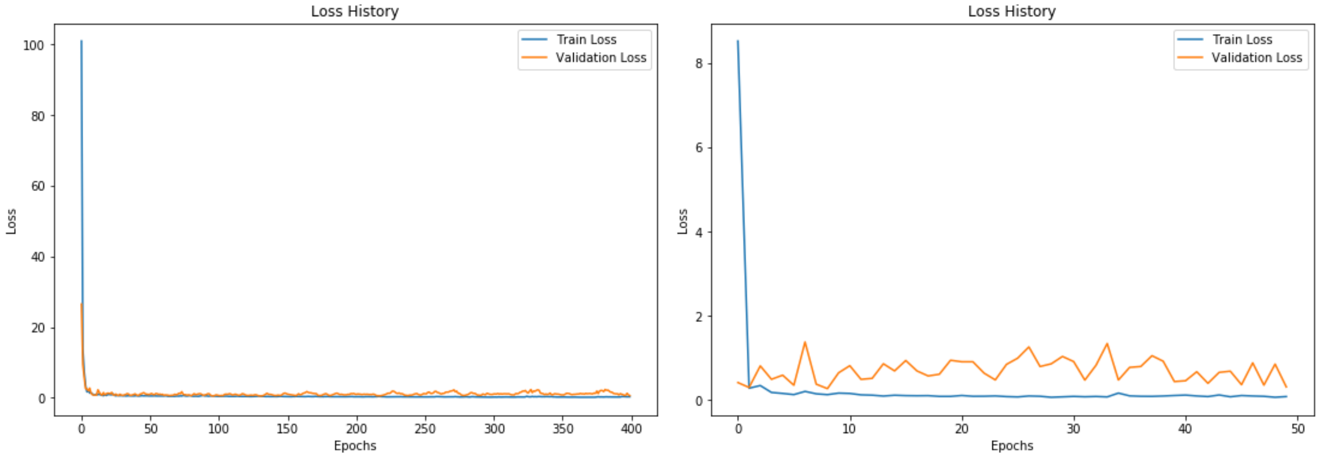


Fig. 3. Train and validation loss for (left) multi-class model and (right) binary-class model.

which we did not replicate, they achieved a test accuracy of 90.7%, with a sensitivity of 88.6% and a specificity of 90.9% [7]. Our final best model results for the first binary model were slightly lower, and exact results were difficult to replicate because the large batch size of 1000 increased randomness across runs. Another factor is due to the data imbalance that we accounted for by under-sampling to the minimum number of images from training, validation, and testing sets. Kermany's test data set consisted of 234 normal images and 390 pneumonia images, while our test data sets had 74 for normal and 148 for pneumonia to account for imbalance.

Many configurations of the model were stuck with roughly 66% accuracy because it was unable to predict any of the viral pneumonia correctly. The 3-class problem likely needed more data to allow the model to learn features that would better differentiate bacterial and viral pneumonia.

While we did see small improvements in performances by using data augmentation, performance could have been increased if we had a greater variety of images since data augmentation could not generate completely new images. For instance, radiologists are provided with supplemental lateral view X-rays when making a diagnosis, however the dataset taken from Kaggle only contained frontal images. The patients health record is also often used as supplemental information that is taken into consideration to make the diagnosis. Predicting diagnoses based on a more comprehensive data set for each patient would be a more appropriate representation of clinical practices and may lead to better performance.

V. Conclusion

We found that convolutional neural networks are effective at diagnosing normal and abnormal chest X-rays with high accuracy and precision, but can struggle with differentiating types of abnormal X-rays. We were able to replicate previous work on the same data set [7]. Future work involves looking at class activation maps to determine what features the model thinks are relevant. These maps could help us interpret the model's decisions, a useful feature in medical diagnosis, and could help determine why it confuses viral with bacterial pneumonia. This project has shown that transfer learning can used to apply large convolutional neural networks to chest X-rays.

Acknowledgment

Thanks to Dr. Paul Sajda PhD, Dr. Andrew Laine PhD, and Arunesh Mittal, for overseeing this project.

References

[1] Ending child deaths from pneumonia and diarrhea. URL: https://www.unicef.org/publications/files/UNICEF-Pneumonia-Diarrhoea-report-2016-web-version5.pdf.

[2] Pneumonia can be prevented - vaccines can help. URL: https://www.cdc.gov/features/pneumonia/index.html.

[3] Pneumonia: Overview, Aug 2018. URL: https://www.ncbi.nlm.nih.gov/books/NBK525774/.

[4] Mohamed A. Elemraid, Michelle Muller, David A. Spencer, Stephen P. Rushton, Russell Gorton, Matthew F. Thomas, Katherine M. Eastham, Fiona Hampton, Andrew R. Gennery, Julia E. Clark, and et al. Accuracy of the interpretation of chest radiographs for the diagnosis of paediatric pneumonia. *PLoS ONE*, 9(8), 2014. http://dx.doi.org/10.1371/journal.pone.0106051 doi:10.1371/journal.pone.0106051.

[5] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115118, 2017. http://dx.doi.org/10.1038/nature21056 doi:10.1038/nature21056.

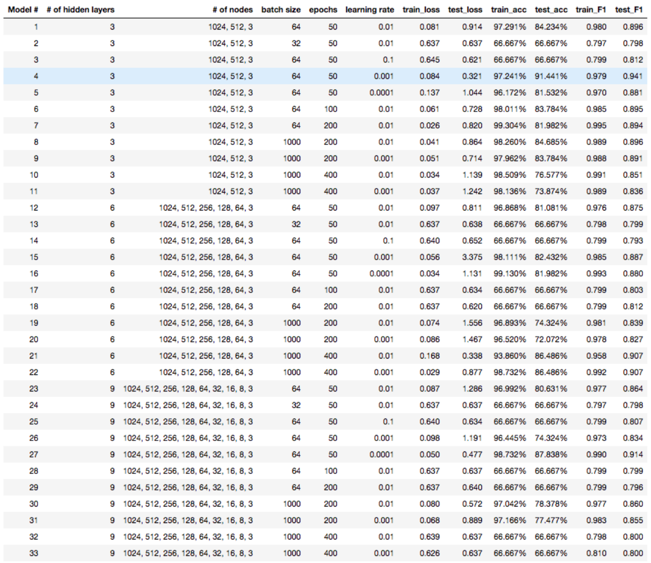
[6] M. Goldbaum, S. Moezzi, A. Taylor, S. Chatterjee, J. Boyd, E. Hunter, and R. Jain. Automated diagnosis and image understanding with object extraction, object classification, and inferencing in retinal images. *Proceedings of 3rd IEEE International Conference on Image Processing*, 1996. http://dx.doi.org/10.1109/icip.1996.560760 doi:10.1109/icip.1996.560760.

[7] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.

[8] Paul Mooney. Chest x-ray images (pneumonia), Mar 2018. URL: https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia.

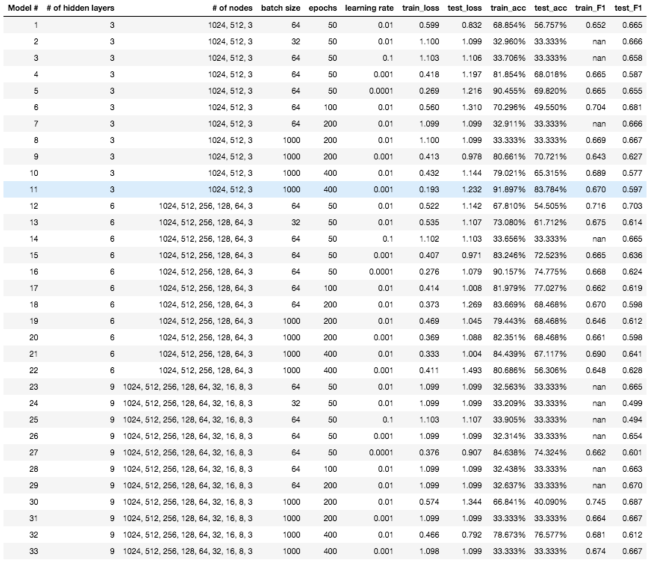
[9] Sangeetha Srinivasan, Sharan Shetty, Viswanathan Natarajan, Tarun Sharma, and Rajiv Raman. Development and validation of a diabetic retinopathy referral algorithm based on single-field fundus photography. *Plos One*, 11(9), 2016. http://dx.doi.org/10.1371/journal.pone.0163108 doi:10.1371/journal.pone.0163108.

VI. Appendix



Model #	# of hidden layers	# of nodes	batch size	epochs	learning rate	train_loss	test_loss	train_acc	test_acc	train_F1	test_F1
1	3	1024, 512, 3	64	50	0.01	0.081	0.914	97.291%	84.234%	0.980	0.896
2	3	1024, 512, 3	32	50	0.01	0.637	0.637	66.667%	66.667%	0.797	0.798
3	3	1024, 512, 3	64	50	0.1	0.645	0.621	66.667%	66.667%	0.799	0.812
4	3	1024, 512, 3	64	50	0.001	0.084	0.321	97.241%	91.441%	0.979	0.941
5	3	1024, 512, 3	64	50	0.0001	0.137	1.044	96.172%	81.532%	0.970	0.881
6	3	1024, 512, 3	64	100	0.01	0.061	0.728	98.011%	83.784%	0.985	0.895
7	3	1024, 512, 3	64	200	0.01	0.026	0.820	99.304%	81.982%	0.995	0.894
8	3	1024, 512, 3	1000	200	0.01	0.041	0.864	98.260%	84.685%	0.989	0.896
9	3	1024, 512, 3	1000	200	0.001	0.051	0.714	97.962%	83.784%	0.988	0.891
10	3	1024, 512, 3	1000	400	0.01	0.034	1.139	98.509%	76.577%	0.991	0.851
11	3	1024, 512, 3	1000	400	0.001	0.037	1.242	98.136%	73.874%	0.989	0.836
12	6	1024, 512, 256, 128, 64, 3	64	50	0.01	0.097	0.811	96.868%	81.081%	0.976	0.875
13	6	1024, 512, 256, 128, 64, 3	32	50	0.01	0.637	0.638	66.667%	66.667%	0.798	0.799
14	6	1024, 512, 256, 128, 64, 3	64	50	0.1	0.640	0.652	66.667%	66.667%	0.799	0.793
15	6	1024, 512, 256, 128, 64, 3	64	50	0.001	0.056	3.375	98.111%	82.432%	0.985	0.887
16	6	1024, 512, 256, 128, 64, 3	64	50	0.0001	0.034	1.131	99.130%	81.982%	0.993	0.880
17	6	1024, 512, 256, 128, 64, 3	64	100	0.01	0.637	0.634	66.667%	66.667%	0.799	0.803
18	6	1024, 512, 256, 128, 64, 3	64	200	0.01	0.637	0.620	66.667%	66.667%	0.799	0.812
19	6	1024, 512, 256, 128, 64, 3	1000	200	0.01	0.074	1.556	96.893%	74.324%	0.981	0.839
20	6	1024, 512, 256, 128, 64, 3	1000	200	0.001	0.086	1.467	96.520%	72.072%	0.978	0.827
21	6	1024, 512, 256, 128, 64, 3	1000	400	0.01	0.168	0.338	93.860%	86.486%	0.958	0.907
22	6	1024, 512, 256, 128, 64, 3	1000	400	0.001	0.029	0.877	98.732%	86.486%	0.992	0.907
23	9	1024, 512, 256, 128, 64, 32, 16, 8, 3	64	50	0.01	0.087	1.286	96.992%	80.631%	0.977	0.864
24	9	1024, 512, 256, 128, 64, 32, 16, 8, 3	32	50	0.01	0.637	0.637	66.667%	66.667%	0.797	0.798
25	9	1024, 512, 256, 128, 64, 32, 16, 8, 3	64	50	0.1	0.640	0.634	66.667%	66.667%	0.799	0.807
26	9	1024, 512, 256, 128, 64, 32, 16, 8, 3	64	50	0.001	0.098	1.191	96.445%	74.324%	0.973	0.834
27	9	1024, 512, 256, 128, 64, 32, 16, 8, 3	64	50	0.0001	0.050	0.477	98.732%	87.838%	0.990	0.914
28	9	1024, 512, 256, 128, 64, 32, 16, 8, 3	64	100	0.01	0.637	0.637	66.667%	66.667%	0.799	0.799
29	9	1024, 512, 256, 128, 64, 32, 16, 8, 3	64	200	0.01	0.637	0.640	66.667%	66.667%	0.799	0.796
30	9	1024, 512, 256, 128, 64, 32, 16, 8, 3	1000	200	0.01	0.080	0.572	97.042%	78.378%	0.977	0.860
31	9	1024, 512, 256, 128, 64, 32, 16, 8, 3	1000	200	0.001	0.068	0.889	97.166%	77.477%	0.983	0.855
32	9	1024, 512, 256, 128, 64, 32, 16, 8, 3	1000	400	0.01	0.639	0.637	66.667%	66.667%	0.798	0.800
33	9	1024, 512, 256, 128, 64, 32, 16, 8, 3	1000	400	0.001	0.626	0.637	66.667%	66.667%	0.810	0.800

Fig. 4. Hyperparameters tested for binary classifier (normal vs. pneumonia). The model with the highest accuracy is highlighted.



Model #	# of hidden layers	# of nodes	batch size	epochs	learning rate	train_loss	test_loss	train_acc	test_acc	train_F1	test_F1
1	3	1024, 512, 3	64	50	0.01	0.599	0.832	68.854%	56.757%	0.652	0.665
2	3	1024, 512, 3	32	50	0.01	1.100	1.099	32.960%	33.333%	nan	0.666
3	3	1024, 512, 3	64	50	0.1	1.103	1.106	33.706%	33.333%	nan	0.658
4	3	1024, 512, 3	64	50	0.001	0.418	1.197	81.854%	68.018%	0.665	0.587
5	3	1024, 512, 3	64	50	0.0001	0.269	1.216	90.455%	69.820%	0.665	0.655
6	3	1024, 512, 3	64	100	0.01	0.560	1.310	70.296%	49.550%	0.704	0.681
7	3	1024, 512, 3	64	200	0.01	1.099	1.099	32.911%	33.333%	nan	0.666
8	3	1024, 512, 3	1000	200	0.01	1.100	1.099	33.333%	33.333%	0.669	0.667
9	3	1024, 512, 3	1000	200	0.001	0.413	0.978	80.661%	70.721%	0.643	0.627
10	3	1024, 512, 3	1000	400	0.01	0.432	1.144	79.021%	65.315%	0.689	0.577
11	3	1024, 512, 3	1000	400	0.001	0.193	1.232	91.897%	83.784%	0.670	0.597
12	6	1024, 512, 256, 128, 64, 3	64	50	0.01	0.522	1.142	67.810%	54.505%	0.716	0.703
13	6	1024, 512, 256, 128, 64, 3	32	50	0.01	0.535	1.107	73.080%	61.712%	0.675	0.614
14	6	1024, 512, 256, 128, 64, 3	64	50	0.1	1.102	1.103	33.656%	33.333%	nan	0.665
15	6	1024, 512, 256, 128, 64, 3	64	50	0.001	0.407	0.971	83.246%	72.523%	0.665	0.636
16	6	1024, 512, 256, 128, 64, 3	64	50	0.0001	0.276	1.079	90.157%	74.775%	0.668	0.624
17	6	1024, 512, 256, 128, 64, 3	64	100	0.01	0.414	1.008	81.979%	77.027%	0.662	0.619
18	6	1024, 512, 256, 128, 64, 3	64	200	0.01	0.373	1.269	83.669%	68.468%	0.670	0.598
19	6	1024, 512, 256, 128, 64, 3	1000	200	0.01	0.469	1.045	79.443%	68.468%	0.646	0.612
20	6	1024, 512, 256, 128, 64, 3	1000	200	0.001	0.369	1.088	82.351%	68.468%	0.661	0.598
21	6	1024, 512, 256, 128, 64, 3	1000	400	0.01	0.333	1.004	84.439%	67.117%	0.690	0.641
22	6	1024, 512, 256, 128, 64, 3	1000	400	0.001	0.411	1.493	80.686%	56.306%	0.648	0.628
23	9	1024, 512, 256, 128, 64, 32, 16, 8, 3	64	50	0.01	1.099	1.099	32.563%	33.333%	nan	0.665
24	9	1024, 512, 256, 128, 64, 32, 16, 8, 3	32	50	0.01	1.099	1.099	33.209%	33.333%	nan	0.499
25	9	1024, 512, 256, 128, 64, 32, 16, 8, 3	64	50	0.1	1.103	1.107	33.905%	33.333%	nan	0.494
26	9	1024, 512, 256, 128, 64, 32, 16, 8, 3	64	50	0.001	1.099	1.099	32.314%	33.333%	nan	0.654
27	9	1024, 512, 256, 128, 64, 32, 16, 8, 3	64	50	0.0001	0.376	0.907	84.638%	74.324%	0.662	0.601
28	9	1024, 512, 256, 128, 64, 32, 16, 8, 3	64	100	0.01	1.099	1.099	32.438%	33.333%	nan	0.663
29	9	1024, 512, 256, 128, 64, 32, 16, 8, 3	64	200	0.01	1.099	1.099	32.637%	33.333%	nan	0.670
30	9	1024, 512, 256, 128, 64, 32, 16, 8, 3	1000	200	0.01	0.574	1.344	66.841%	40.090%	0.745	0.687
31	9	1024, 512, 256, 128, 64, 32, 16, 8, 3	1000	200	0.001	1.099	1.099	33.333%	33.333%	0.664	0.667
32	9	1024, 512, 256, 128, 64, 32, 16, 8, 3	1000	400	0.01	0.466	0.792	78.673%	76.577%	0.681	0.612
33	9	1024, 512, 256, 128, 64, 32, 16, 8, 3	1000	400	0.001	1.098	1.099	33.333%	33.333%	0.674	0.667

Fig. 5. Hyperparameters tested for multi-classifier (normal vs. bacterial vs. viral pneumonia). The model with the highest accuracy is highlighted.