JOURNAL OF
EMERGING INVESTIGATORS

Article

Determining the best convolutional neural network for identifying tuberculosis and pneumonia in chest x-rays

Sunny Kudum[1,2], Dristi Shah[3], Ishita Vaish[4], Padmavati Tirumala[5]

[1] Academy of Science, Leesburg, Virginia
[2] Riverside High School, Leesburg, Virginia
[3] Rutgers Preparatory School, Somerset, New Jersey
[4] Northview High School, Duluth, Georgia
[5] IBM, 2300 Dulles Station Blvd, Herndon, Virginia

SUMMARY

Tuberculosis (TB) and pneumonia are commonly misdiagnosed respiratory conditions associated with high rates of mortality. Chest X-rays (CXRs) are an inexpensive method to identify respiratory conditions. Thus, a model used to distinguish between CXRs depicting lungs classified as normal, pneumonia, and tuberculosis would lead to accurate diagnoses of these respiratory conditions. This need is fulfilled by the recent explosion in deep learning, and new models with robust performance are constantly developed. However, these models have varying strengths and weaknesses which allow them to excel at certain tasks and struggle with others. Therefore, testing these models is essential to find the most suited model. In this study, we trained and applied six convolutional neural networks, the InceptionV3, ResNet50, ResNet152, InceptionResNetV2, DenseNet121, and AlexNet, to the diagnosis of TB and pneumonia. We hypothesized that InceptionResNetV2 would perform best for this task due to its combination of inception blocks that reduce the dimensionality of the CXR images and residual blocks that allow for deeper models by eliminating vanishing gradient. After training on a combination of five datasets from the Guangzhou Women and Children's Medical Center, Shenzhen, Montgomery County, Belarus, and ChestX-ray8, it was found that various models excelled in predicting different diseases shown in the datasets. The results displayed that there was no clear superior model but instead significant superiority within certain diseases.

INTRODUCTION

In 2018, over 10 million people fell ill with tuberculosis (TB), with a majority of these cases concentrated in the developing countries of India, China, Indonesia, the Philippines, Pakistan, Nigeria, Bangladesh, and South Africa (1). Tuberculosis is an infectious disease caused by *Mycobacterium tuberculosis* (MTB), generally targeting an individual's respiratory system. When infected, MTB multiplies in its host's lungs, destroys lung tissue, and eventually spreads to other parts of the body through the bloodstream or lymphatic system (2). Another challenge posed by TB is the difficulty in distinguishing between TB and other diseases, such as pneumonia. Pneumonia causes an individual's pulmonary air sacs to fill up with fluid or pus (3). Pneumococcal pneumonia, the most common type of bacterial pneumonia, is a deadly disease that typically affects one lobe of the lung and can develop following an instance of the flu or a common cold (3). This disease spans an even wider range than TB as a common respiratory condition that affects over 450 million people every year (1).

The similarities between chest-x-ray (CXR) scans of tuberculosis and pneumonia are one of the leading causes of misdiagnoses. This is largely due to the lack of specialized faculty, primarily in developing countries, that can differentiate between these two diseases as well as the similarity in clinical and radiological patterns of TB and pneumonia (4). Although CXRs are an inexpensive and rapid method used to identify lung abnormalities by portraying complications both within and around the lung—thus conventionally used to diagnose a variety of pulmonary conditions (5)—misdiagnosis remains a pressing issue. This is because, in the acute phase, tuberculosis and pneumonia look extremely similar, which can be a particularly damaging problem in areas with a lack of trained radiologists. Therefore, automating these diagnoses through deep learning models trained on CXRs offers a promising avenue to better differentiate between pneumonia and TB. The increasing availability of datasets enhances the ability of models to identify features that can conclude the presence or lack of a respiratory condition. Additionally, the use of models trained on multi-institutional datasets reduces bias generated from the circumstances where the CXRs were taken, making the deep learning models robust.

To predict disease states based on CXRs, this study utilizes convolutional neural networks (CNNs) that help enhance image classification of the x-rays by understanding patterns in x-rays displaying each disease. CNNs are deep learning algorithms inspired by the human visual cortex. They are currently the most popular technique for image classification in the biomedical field and commonly consist of different layers that provide the prediction power. They can consist of convolutional layers, pooling layers, fully connected layers, and many other sophisticated aspects (6). Convolutional layers are filters that are utilized to transform the image and pass the results to the next layers, whereas a pooling layer is normally utilized to reduce the feature map of the image to focus on important details. This is often achieved by identifying the average presence of a feature (average pooling) or by finding the most often activated presence in the image. Combinations of these layers (and other more complicated structures) allow for the construction of models that can be used to make predictions on image data. Training these powerful models requires several runs through the training data, and each run is known as an epoch. The vast array of combinations that can be used to make models has resulted in some extremely successful models for prediction tasks. Several of the most powerful deep learning techniques in the classification of respiratory conditions with CXRs include AlexNet (7), ResNet50 (8), ResNet152 (8), InceptionV3 (9), InceptionResNetV2 (9), and DenseNet121 models (10). This study compares the efficacy of these models to differentiate between CXRs of patients with pneumonia and tuberculosis to provide a tool for radiologists when diagnosing patients. We hypothesized that the InceptionResNetV2 would be the best performing model to distinguish tuberculosis and pneumonia from CXRs but in reality, there was similarity in the overall performance of the models. The differences were stark when analyzing the models' performance on individual disease states (tuberculosis, pneumonia, and normal).

RESULTS

The world of artificial intelligence is rapidly growing, and along with it, is the variety of CNNs. These CNNs have differing architectures and world-class performance but are not a one-size-fits-all solution.

To determine what is best for this vital problem, we aim to use acclaimed models and form conclusions on what tools radiologists would benefit most from using.

Data for training was derived in the form of images provided by Guangzhou Women and Children's Medical Center, Shenzhen, Montgomery County, Belarus, and ChestX-Ray8.

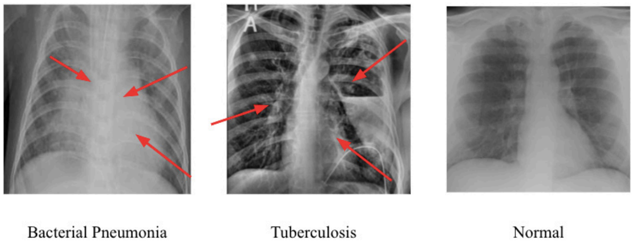


Figure 1: Images of Tuberculosis, Bacteria Pneumonia, and Normal Chest X-Rays (representation of all three classes in the assembled dataset). Three sample images of different disease states picked randomly from our combined dataset. The red arrows on the image point to telltale signs of the abnormalities that come with the disease that classifies it.

We combined these different datasets and images into one large dataset (**Table 1**) with three total classes of x-ray images (**Figure 1**).

We then preprocessed this combined dataset into training and validation sets, resized, and augmented (**Figure 2**). Using Keras libraries, we constructed the models and evaluated them with metrics of accuracy, specificity, recall, precision, loss, and F1 score. In this work, precision describes the ability of the model to predict every positive value as positive. "Positive" samples consist of any disease/condition that is in question in the confusion matrix. The recall value provides the fraction of positive samples correctly predicted by the model as positive. Therefore, this value provides the consistency with which a particular model can predict the class each time it is shown in the dataset. Specificity is similar as it provides

	Guangzhou Women and Children's Medical Center	Shenzhen	Belarus	Montgomery County	ChestX-ray8	Total
Tuberculosis	N/A	336	304	58	N/A	698 (Training: 558, Validation: 140)
Bacterial Pneumonia	2804	326	N/A	N/A	N/A	3130 (Training: 2504, Validation: 626)
Normal	1583	N/A	N/A	80	864	2527 (Training: 2022, Validation: 505)

Table 1: Data Breakdown Based on Disease and Dataset. The amount of CXRs in the different datasets used to train the models. They are further broken down by the amount of CXRs in each class within each dataset.

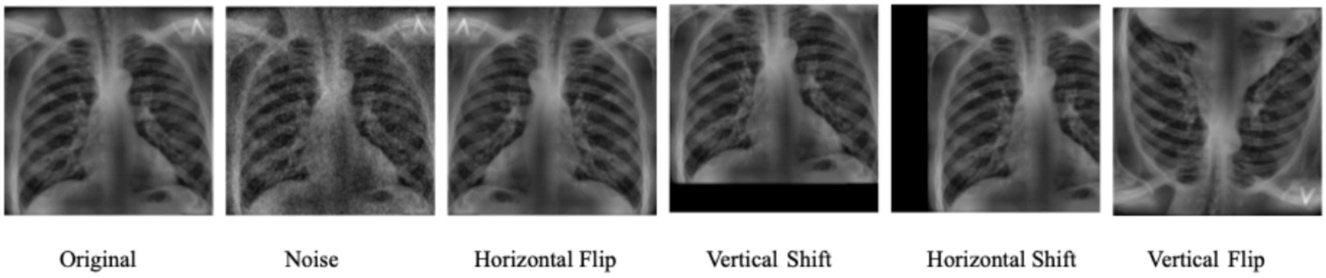


Figure 2: A sample image under each augmentation performed in the experiment. Examples of the augmentations that were performed on the image (the type of augmentation is shown below each image). Augmentations relevant to possible mishaps in the process of administering chest x-rays were included to increase relevance to prediction.

the fraction of negative samples correctly predicted by the model as negative. For this multiclass problem, negative classes are denoted as any class other than the one being currently presented to the model. For example, if calculating the specificity for tuberculosis, a negative class will be either bacterial pneumonia or normal conditions. The F1 score provides a depiction of the balance between the precision and recall metrics for each model and accounts for the class imbalance within the dataset.

After training was completed and these metrics were outputted, we generated graphs of validation accuracy and loss over the epochs (**Figures 3, 4)** as well as confusion matrices for all six models (**Figures 5, 6**). In these figures, we found that many of the models presented similar metrics and that they cannot be differentiated significantly with just an overall look at their metrics as all metrics were greater than 0.9 (**Table 2**).

We see a similar trend with precision, recall, and F1 score, which all have a difference of 0.2 between the lowest and highest performing models on each metric. The specificity values of all six models are 0.97. Nevertheless, the values we see here are still important to the overall analysis of the models. The InceptionV3 and DenseNet121 achieved higher precision values of 0.95, indicating a low-false positive rate. The InceptionV3, along with the ResNet50, also had the highest recall values of 0.95 and 0.94 respectively, indicating a low rate of incorrect prediction on the negative class. The DenseNet121 and InceptionV3 both yielded an F1 score of 0.95, suggesting higher overall precision and recall scores.

To account for the overall similarity of these metrics, we plotted the validation accuracy and loss of each model over the epochs (the number of times the model goes through the training set) and noted how long models ran before early stopping (a method by which the model is stopped from overfitting on the training set by measuring if the accuracy/loss is improving through epochs). With these graphs, we can see not only the final results of the model, but also how they fared during their training and if their metrics are a result of overfitting. The DenseNet121 and InceptionResNetV2 had relatively less spikes and maintained a healthy pattern of metrics throughout all epochs prior to early stopping when compared to other models like the ResNet152 or InceptionV3. We also analyzed each model's performance on the three

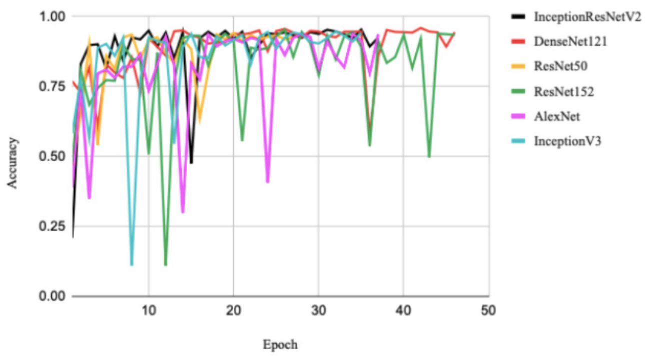


Figure 3: The validation accuracy across epochs for each model. The validation accuracy for all models (the key is given on the right of the graph) over the epochs that they trained for. The individual lines stop at different areas due to when its training stopped due to early stopping.

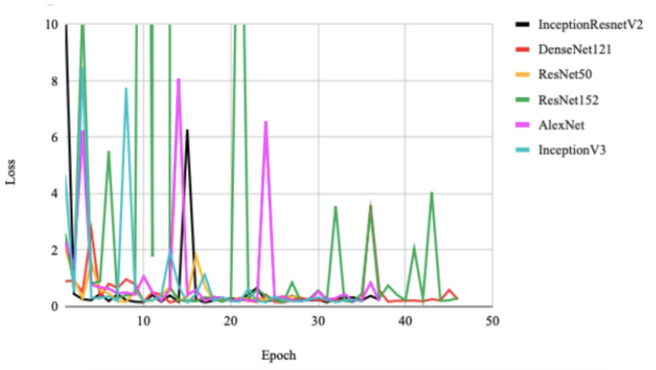
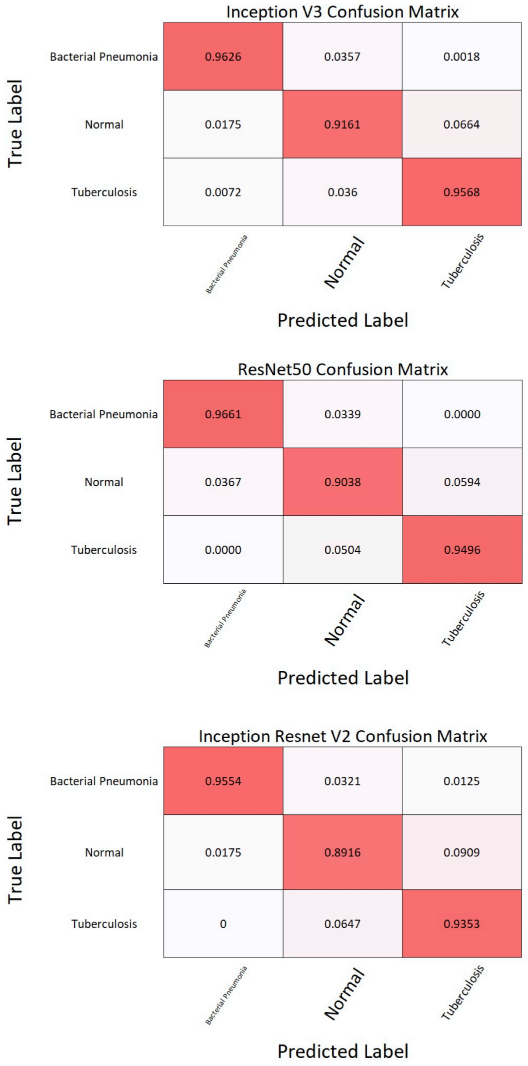


Figure 4: The validation loss across epochs for each model. The validation loss for all models (the key is given on the right of the graph) over the epochs that they trained for. The individual lines stop at different areas due to when its training stopped due to early stopping.



Inception V3 Confusion Matrix

	Bacterial Pneumonia	Normal	Tuberculosis
Bacterial Pneumonia	0.9626	0.0357	0.0018
Normal	0.0175	0.9161	0.0664
Tuberculosis	0.0072	0.036	0.9568

ResNet50 Confusion Matrix

	Bacterial Pneumonia	Normal	Tuberculosis
Bacterial Pneumonia	0.9661	0.0339	0.0000
Normal	0.0367	0.9038	0.0594
Tuberculosis	0.0000	0.0504	0.9496

Inception Resnet V2 Confusion Matrix

	Bacterial Pneumonia	Normal	Tuberculosis
Bacterial Pneumonia	0.9554	0.0321	0.0125
Normal	0.0175	0.8916	0.0909
Tuberculosis	0	0.0647	0.9353

Figure 5: Confusion Matrices for the Inception V3, InceptionResnetV2, and the ResNet50. For these confusion matrices depicting the performance of the Inception V3, InceptionResnetV2, and the ResNet50, the intensity of the color corresponds to the normalized accuracy of each square in the matrix.

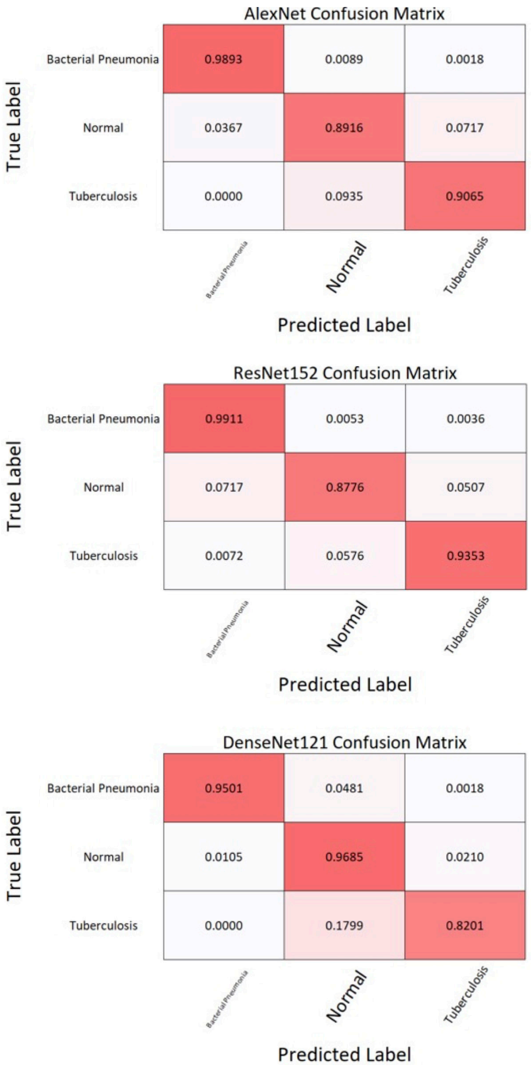


Figure 6: Confusion Matrices for the AlexNet, ResNet152, and DenseNet121. For these confusion matrices depicting the performance of the AlexNet, ResNet152, and DenseNet121, the intensity of the color corresponds to the normalized accuracy of each square in the matrix.

classes using confusion matrices (CMs) (**Figures 5, 6**). Confusion matrices are tables that allow for easy visualization of model performance. The CMs depict nine values, which represent the normalized accuracy in each class and are placed on a color scaled based on this accuracy for each class. The AlexNet and ResNet152 models performed relatively well in classifying bacterial pneumonia CXRs (accuracies of 98.93% and 99.11%, respectively) when compared to the other four models. The AlexNet, ResNet152, and InceptionResNetV2 models performed worse in the normal class. By a small margin, the ResNet50 and InceptionV3 models performed better while classifying normal classes. The DenseNet121 showed great proficiency in classifying normal CXRs with a 96.85% accuracy but also performed the worst on tuberculosis images (82.01% accuracy). This sort of polarity in models suggests different strengths and weaknesses and therefore, different models would fit best to respective purposes.

DISCUSSION

Tuberculosis and pneumonia are both life-threatening respiratory conditions that disproportionately affect those residing in developing countries. Because these countries may lack the medical personnel needed to quickly diagnose TB and pneumonia, it is critical to maximize the potential of deep learning on CXR-based diagnosis. Furthermore, the WHO has consistently elaborated on the necessity of applying deep learning models on CXRs to prevent misdiagnosis of common respiratory conditions. This study addressed this need by comparing different deep learning models trained on datasets from five different sources to distinguish TB and pneumonia. Convolutional neural networks trained on multi-institutional datasets may aid radiologists in delivering diagnoses more accurately and quickly.

We hypothesized that the InceptionResNetV2 would be the best performing model to distinguish tuberculosis and pneumonia from CXRs. In reality, however, there was a high level of similarity among the tested CNN's. We depicted the metrics in only two and four significant figures (in the case of accuracy) because there is little difference in the overall metrics. Similarly, all models showed similar noise levels, but on this slim margin, the DenseNet121 was the best performing model in terms of average validation accuracy and validation loss. DenseNet121 may have outperformed

	Precision	Recall	F1 Score	Accuracy	Specificity
IR2*	.94	.93	.93	.9245	.97
DenseNet121	.95	.91	.95	.9442	.97
ResNet50	.94	.94	.94	.9363	.97
AlexNet	.94	.93	.94	.9363	.97
ResNet152	.94	.93	.93	.9340	.97
InceptionV3	.95	.95	.95	.9410	.97

*IR2 denotes Inception-ResNetV2

Table 2: Classification Metrics for Trained Models on Validation Data. The precision, recall, F1 score, accuracy, and specificity for each model on validation data. The accuracy values were within 0.2 percentage of each other (all around 0.93 and 0.94).

the other models due to its concatenating feature maps, which allow for maximum information flow between layers. However, as previously highlighted, there was no real superior model in this study. The similarities between the metrics achieved by the models made the use of further statistical analysis unnecessary as results would not show any significant difference (**Table 2**). Instead, the confusion matrices demonstrated that each model has strengths and weaknesses in different aspects of the prediction (**Figure 5**). The validation accuracy and loss showed variation and inconsistency, which can mostly be attributed to the use of the mini batch gradient descent with the Adam optimizer (chance could result in unlucky iterations) and a learning rate that is too large. Furthermore, the integration of multiple datasets could result in noise in the data, which has not been addressed in this work in order to get a better comparison of how the tested models can handle this noise. Some models, like the InceptionResNetV2 and DenseNet121, were able to handle the noise, Adam optimizer, and large learning rate better than others.

The performance of the six models on the three classes of data (tuberculosis, bacterial pneumonia, and normal) was compared using confusion matrices. When comparing the confusion matrices to select the best performing model, it was clear that each model was better at performing different tasks. While the bacterial pneumonia class consistently yielded the highest accuracy, it varied slightly between the models (**Figure 5**). While we assumed that the models would struggle with distinguishing between TB and bacterial pneumonia, the confusion matrices show that the models instead struggled to distinguish these diseases from the normal class. This may be because the 'normal' class consisted of any x-rays that did not depict TB, bacterial pneumonia, or any of the 14 most common thoracic diseases. This is a broad category that may have x-rays showing various other diseases that the models may struggle to form accurate decision boundaries. Furthermore, the multi-institutional nature of the data used could result in further confusion for the models, as different sources of data could provide various angles or quality of pictures. Although this could result in a lower perceived accuracy, the use of the multi-institutional aspect improves the robustness and practical usage of the model. The models that did well while predicting TB (InceptionV3 and ResNet50) show a greater ability to generate accurate classifications with a lower number of real images (not augmented) than the other models. This may be because they have a shallower structure than that of the ResNet152 and InceptionResNetV2, allowing them to learn the conditions of a pattern easier.

While there are evident differences between the models, they all achieve high metrics, meaning that there is no model that would bring a detriment to the overall prediction task. This relative similarity with finer differences shows that our hypothesis was proven incorrect. While we hypothesized that the InceptionResNetV2 would perform the best for the prediction of tuberculosis, pneumonia, and normal lungs, this model's performance varied based on the class, and other models excelled in classes the InceptionResNetV2 struggled in.

The majority of the models in this study were first built and trained on the ImageNet, where their differences in architecture produced significant difference in performance. However, the ImageNet dataset is much larger and complex than our dataset, which focuses solely on Chest X-rays. This could provide less room for the architecture to form different patterns, and therefore, different results, which leads to the conclusion formed in this study.

Some limitations of our study included a lack of original TB CXRs. The datasets we used to provide TB images did not provide as many images as those used in the bacterial pneumonia class. As a result, we performed more augmentations on the tuberculosis class to even the class imbalance. While this is an acceptable solution as the model learns to identify TB despite the disturbance, more TB CXRs would be optimal. Another limitation includes that there was a lack of adult CXRs to train the bacterial pneumonia class as the Guangzhou Women and Children's Medical Center dataset had CXRs for only those from one to five years old. This dataset comprises much of the bacterial pneumonia class, so the changes in the chest that occur as patients age may not be accounted for. Furthermore, the metrics shown in this study may not be completely accurate with fine tuning with respect to each model. With the purpose of not interfering with the metrics achieved by each model, we did not fine tune the models based on their weaknesses and strengths (after all, these are the weaknesses and strengths that we sought to identify). However, we still provide the baseline reality for the performance of these models in this experiment.

The scope of this research can be expanded by incorporating more diseases that are misdiagnosed such as lung cancer (11). To improve upon the accuracy of the constructed models, transfer learning could be applied. Transfer learning is a technique by which a machine learning a model trained for one is re-used and changed for utilization on another purpose. It could be utilized here to improve the models in this study with patterns learned from application to another topic (12). Also, because this study sought to manipulate a limited number of variables, there was little fine-tuning on each model. Therefore, testing how optimal performance varies with a change in hyperparameters may provide insight into the application of neural networks to label CXRs. The discrepancies between the models' performance in each class show the various approaches and advantages of the six models. The differing architectures and abilities of the models uniquely influenced their decision-making for the three classes of CXRs in this study. In conclusion, these results provide a general overview of how these models can be used in future applications.

METHODS AND MATERIALS

To test our models, we first had to get our data. To minimize bias, we combined data from five different datasets that summarized three classes: tuberculosis, bacterial pneumonia, and normal. Our tuberculosis datasets (Shenzhen, Belarus, Montgomery County) contained different varieties of tuberculosis manifestations, which we noted within the dataset. Some images in the data depicted spinal tuberculosis (STB) while others showed pulmonary tuberculosis (PTB) and bilateral pulmonary tuberculosis. The region of this disease was also identified in the description of the images. In contrast, the pneumonia datasets we used did not identify the specific type of pneumonia (lobar, multifocal, etc.). However, they did have classifications of viral and bacterial pneumonia (only bacterial pneumonia was used for this experiment).

CXR images from the Shenzhen Dataset included 326 normal CXRs (defined as not containing CXRs which depict TB, pneumonia, or any of the fourteen most common thoracic diseases) (13) and 336 abnormal CXRs with various manifestations of tuberculosis (including spinal tuberculosis, pulmonary tuberculosis, and bilateral pulmonary tuberculosis). The Montgomery Dataset from the Department of Health and Human Services of Montgomery County contained 80 normal CXRs and 58 TB CXRs (14). This dataset included effusions, military patterns, and other abnormalities. The Guangzhou Women and Children's Medical Center dataset (15) included images of CXRs depicting normal lungs, bacterial pneumonia, or viral pneumonia; however, we only utilized CXRs of normal lungs and lungs with bacterial pneumonia, totaling 4387 images. This dataset had a large range of pneumonia patients, but the specific type of pneumonia (lobar, multifocal, etc.) was not noted. The Belarus (16) dataset consisted of 304 CXRs, all of which showed patients infected with TB. From the ChestX-ray8 dataset published by the National Institutes of Health, we extracted 864 CXRs of normal lungs. We incorporated datasets with CXRs of bacterial pneumonia, TB, and normal lungs (**Figure 1**) to improve the likelihood that the model can accurately differentiate between various conditions.

We then combined these images into one dataset and divided them into training and validation sets with an 80:20 ratio, respectively (**Table 1**). We resized all images in the dataset to 224x224 pixels. The data was augmented to generate new, unique images to create an equal number of images in each class. In this work, we used augmentation for a dual purpose. It helped create a more robust model that can identify diseases despite added noise/alterations to the images while also fixing the class imbalance. We implemented the typical augmentation methods, which consist of a horizontal flip, vertical flip, noise addition, image blur, width shift, and height shift, on all images in the training set. Before the augmentations, there were 5079 lung CXRs in the training set. Data in the training sets were augmented to form a total of 10,500 images, with each class of tuberculosis, bacterial pneumonia, and normal containing 3,500 CXRs. However, we did not augment the validation set to reduce bias during the evaluation of the model's performance and consisted of 1272 images.

With the images ready, we then moved on model construction, which we did through understanding and building the architecture of our acclaimed models with Keras libraries. These models were set to predict three classes with an input shape of 224x224x3 (the image dimensions that we reshaped to). After each model was instantiated, we compiled them with an Adam optimizer (an optimization algorithm that updates weights based on training data) and set to output metrics of accuracy, specificity, recall, precision, loss, and F1 Score. We then trained each model on the previously preprocessed data and set them to run for 50 epochs (early stopping and best model callbacks were declared to prevent overfitting).

Utilizing these metrics (which were now stored in arrays for every epoch that the model underwent), we formed graphs for each metric. Furthermore, the classification report and confusion matrix toolkits were utilized to receive information about the performance of the models. Using a formatting method, these were then turned into interpretable confusion matrices for each model. The models that we used in this experiment are explained below.

AlexNet

AlexNet revolutionized deep learning in 2010. It produced breakthrough results in the ImageNet LSVRC-2010 contest, achieving a top-1 error rate of 39.7%. The AlexN*et al*gorithm consists of five convolutional layers and three fully connected layers. Multi Convolutional Kernels, part of the convolutional blocks in the AlexNet, extract features from the image. The AlexNet was pre-trained on 1.3 million images, validated on 50,000, and tested on 150,000 (7).

Residual Networks

Residual Neural Networks (ResNets) are a powerful set of algorithms used for image classification. Residual networks address the vanishing gradient problem, in which accuracy decreases as the number of layers increases. Instead of stacking convolution layers, ResNets add a skip connection, which adds a convolutional block to the input and output. This mitigates the problem by allowing an alternate shortcut path for the gradient to flow through. Both the ResNet50 and the ResNet152 models were used (8).

InceptionResNetV2

The InceptionResNetV2 is based upon a combination of the structure of the Inception and ResNet models. The most basic architecture of this model consists of a stem block, five repetitions of the Inception-ResNet-A, a reduction, ten repetitions of the Inception-ResNet-B, a reduction, five repetitions of Inception-ResNet-C, average pooling, dropout, and then a softmax activation function. Outputs are concatenated before each inception module (9).

InceptionV3

The InceptionV3 was created to minimize representational bottleneck, which refers to the loss of information when convolutions alter the input's dimensions. Thus, a 5x5 convolution is factored into two 3x3 convolution operations to increase performance and computational speed. Instead of increasing depth, the InceptionV3 widens filter banks. It builds upon the InceptionV2 through four primary additions: an RMSProp Optimizer, Factorized 7x7 convolutions, BatchNorm in the Auxiliary Classifiers, and Label Smooth Regularization. In particular, label smooth regularization estimates the dropout rate to decrease the error rate. The InceptionV3 model constructed in this study consisted of a raw InceptionV3: average pooling, max pooling, concatenation, dropout, a fully connected Dense layer, and a concluding SoftMax function (9).

DenseNet

DenseNets, or Dense Convolutional Networks, were created to provide an alternative way to increase the depth of deep convolutional neural networks. Classic neural networks connect the output of one layer to the next after performing operations. While other networks sum the output feature maps of a layer with the incoming feature maps, DenseNets concatenate them.

The output of the previous layer acts as an input for the second layer by using composite function operation (which consists of the convolution layer, pooling layer, batch normalization, and non-linear activation layer). This study utilizes the DenseNet121, which consists of five convolution and pooling layers, three transition layers, one classification layer, and two dense blocks. The DenseNet121 allows for the use of deeper models without concern for vanishing gradients (10).

ACKNOWLEDGMENTS

We thank our colleagues Rahul Kumar, Varun Pasupuleti, and Kamran Majid from the Aluna Research Group who provided insight and expertise that greatly assisted the research process.

Received: February 28, 2021
Accepted: August 17, 2021
Published: February 9, 2022

REFERENCES

1. Global tuberculosis report 2018. Geneva: World Health Organization; 2018. Licence: CC BY-NC-SA 3.0 IGO
2. Zaman, Kalequ. "Tuberculosis: a global health problem." *Journal of Health, Population, and Nutrition*, vol. 28, num. 2, 2010, pp. 111-113.
3. "What is Pneumonia?" *American Journal of Respiratory and Critical Care Medicine*, vol. 193, num. 1, 2016, pp. P1-P2
4. Laushkina, Zhanna. "The Analysis of Factors Associated with Misdiagnosis Pneumonia in TB Hospital." *European Respiratory Journal*, vol. 46, 2015, pp. PA1524.
5. Singh, Ramandeep, *et al*. "Deep Learning in Chest Radiography: Detection of Findings and Presence of Change." *PLOS ONE*, vol. 13, no. 10, 2018, pp. e0204155.
6. Lee, June-Goo, *et al*. "Deep Learning in Medical Imaging: General Overview." *Korean Journal of Radiology*, vol. 18, no. 4, 2017, pp. 570-584, doi:10.3348/kjr.2017.18.4.570.
7. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton "ImageNet Classification with Deep Convolutional Neural Networks." *Advances in Neural Information Processing Systems*, vol. 25, num. 2, 2012, pp. 1097–1105.
8. He, Kaiming, *et al*. "Deep Residual Learning for Image Recognition." *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, doi:10.1109/cvpr.2016.90.
9. Raj, B. "A Simple Guide to the Versions of the Inception Network Medium." *Towards Data Science*, 30 May 2018, https://towardsdatascience.com/a-simple-guide-to-the-versions-of-the-inception-network-7fc52b863202.
10. Huang, Gao, *et al*. "Densely Connected Convolutional Networks." *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. http://arxiv.org/abs/1608.06993.
11. National Organization for Rare Diseases. "Nontuberculous Mycobacterial lung disease." *NNORD Rare Disease Database*, Jan 2 2019. https://rarediseases.org/rare-diseases/nontuberculous-mycobacterial-lung-disease/
12. Tan, Chuanqi, *et al*. "A Survey on Deep Transfer Learning." *The 27th International Conference on Artificial Neural Networks (ICANN)*, 2018, http://arxiv.org/abs/1808.01974(17)
13. Wang, Xiaosong, *et al*. "ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases." *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, doi:10.1109/cvpr.2017.369.
14. Jaeger, Stefan, *et al*. "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases." *Quant Imaginig Med Surg*, vol. 4, num 6., 2014, pp. 475-477.
15. Kermany, Daniel, Kang Zhang, and Michael Goldbaum. "Identifying medical diagnoses and treatable diseases by image-based deep learning." *Cell*, vol. 172, num. 5, 2018, pp. 1122-1131.
16. B.P. Health. "Belarus Tuberculosis Portal." https://www.tuberculosis.by/.