Prediction of Tuberculosis and Pneumonia from Chest X-ray Images

Abstract

Pneumonia and tuberculosis kill four million people each year, making them one of the leading causes of death worldwide. One of the methods used to diagnose these infections is a chest X-ray. Major deaths occur in developing countries However, it is extremely difficult for a well-trained radiologist to examine the chest X-rays to distinguish the infections. Hence more efficient ways are required for the diagnosis of these infections. In this project, we propose an efficient model for the classification of these infections and provide aid to radiologists. Various models were configured and tested. A supervised learning approach where CNN, VGG16, and Random Forest Classifier were used in order to train the models. Another approach composed of feature extraction where the convolution neural networks technique was used and then passed through traditional neural networks. These combinations of neural networks and machine learning algorithms provided a high accuracy and recall. The VGG16 + Random Forest model provides an accuracy of 98% and a recall score of 98%. Thus, this combined model can be used for the diagnosis of pneumonia and tuberculosis and assist the radiologist

Keywords: Pneumonia, Tuberculosis, VGG, Random Forest, Convolution Neural Network, Recall, Precision, Accuracy.

1. Introduction:

Pneumonia is a common disease that is also a leading cause of death worldwide. It is a severe respiratory illness caused by viruses and bacteria that affects people of all ages. It is a lung disease caused by inflammation of the lung tissue caused by the pathogen filling up the alveoli with pus or fluid, reducing the exchange of carbon dioxide and oxygen between the blood and the lungs. Fever, shortness of breath, chest pain, and cough are the most common symptoms. Children under the age of five, adults over 65, and people with preexisting health problems are the most vulnerable. In 2017, an estimated 2.56 million people died due to pneumonia in the world. Almost one-third of the victims i.e., 800,000 were children below the age of 5. Pneumonia is most observed in South Asia and the Sub-Saharan region of Africa.

Tuberculosis is another leading cause of death worldwide. It is a bacterial infection that most commonly affects the lungs, but in rare cases, other organs can be affected as well. The bacteria colonize the lungs and destroy lung tissue, causing the victim to cough and spread the virus through the air. The most common symptoms are chest pain, fever, weakness, and a persistent cough. Most people affected are adults in their productive years. According to the World Health Organization, tuberculosis sickens 10 million people each year and kills 1.5 million. Approximately 95% of these cases are discovered in developing countries.

One of the methods for determining pneumonia and tuberculosis is a chest X-ray. Another test is a CT scan, which produces high-resolution images. However, X-rays are preferred over CT scans because the images take longer to produce, and many developing countries may not have high-resolution CT scanners. There are several regions around the world where there is a shortage of healthcare workers and radiologists whose prediction of such diseases is important. In many developing countries, clinical officers lack the necessary training, and the symptoms of the disease are frequently confused with those of other diseases. As a result of these factors, the radiologist has difficulty diagnosing and predicting the correct disease.

Prediction using computer-aided artificial intelligence is becoming increasingly popular. This service can be provided to a large population at a low cost. Deep neural networks and machine learning outperform the average radiologist in disease prediction accuracy, and these methods address the issues that plague radiologists in developing countries.

2. Related Work:

For many years, the detection of pneumonia using chest X-rays has been an open problem, with the main limitation being a lack of publicly available data. Traditional machine learning methods have been extensively researched.

We referred to a paper proposed by Albahli S [6] in which simple CNN architectures were developed for the classification of pneumonic chest X-ray images. They used data augmentation to compensate for the scarcity of data, and the accuracy rates obtained were 90 percent and 93 percent, respectively. Whereas on other hand, they used the DenseNet-121 CNN model [21] for pneumonia classification, but only got a 76.8 percent f1-score. They suspected that the unavailability of patient history was a major cause for the inferior performance of both their deep learning model and the radiologists with which they compared the performance of their method.

As a result, we devised a combination of deep neural networks and traditional ML algorithms to address these issues.

3. **Background:** In response to the above situation, Pasa, F., Golkov [2] proposed a CNN method, which could automatically extract features through continuous layers and output the possibility of which class the input images belonged to. The essence of CNN is to filter previous images or feature maps through a specific convolution kernel to generate the feature map of the next layer, and then combine it with operations such as pooling operations to reduce feature map scale and computation. The generated feature map is then enhanced with a nonlinear activation function to improve the model's characterization ability. Maximum pooling and average pooling are two common pooling operations. Maximum pooling means that the feature delivered into the pooling layer is divided into sub-regions and will output the maximum of each sub-region based on the horizontal and vertical strides. The only difference between the maximum and average pooling is the output of the sub-region where the pooling occurs. Where pooling outputs the average of each sub-region.

With the extensive development of deep learning in recent decades, the most popular neural framework has been proposed. AlexNet and VGGNet [3] are two examples. However, when the number of network layers is increased, As the number of neurons in the neural network grows, the problem of gradient disappearance or gradient explosion becomes fixated on specific features of the training image rather than learning more generative features, limiting the model's generalizability. Ability suffers as a result of overfitting, here the functions include ReLU (Rectified Linear Units) and Sigmoid.

In this study, we have used the traditional method of Random Forest along with the VGG16 in order to improve the accuracy and recall rates.

4. Problem definition:

Chest X-rays and CT scans are the methods used to determine both pneumonia and tuberculosis, but as it takes more time to produce the images and scarcity of healthcare workers in underdeveloped countries, radiologists are switching to AI.

We present the detailed experiments and evaluation steps undertaken to test the effectiveness of the proposed model. Our experiments were based on a chest X-ray image dataset. We deployed Kera's opensource deep learning framework with the TensorFlow backend to build and train the convolutional neural network model and then used VGG 16 for tuning hyperparameters and a random forest classifier for implementing the recall function.

5. Implementation:

Three different models for the classification of chest X-ray images were investigated in this project.

Data: -

The dataset we used in our project consists of X-ray images that were taken from the different publicly available datasets like

1. National Library of Medicine Dataset.

2. Belarus Dataset.

3. NIAID TB Dataset.

4. RSNA Dataset.

5. Mendeley Dataset.

In total, the data set consists of 3,734 normal X-ray images, 3836 TB infected images, and 4125 pneumonia infected images combining all the images from the above publicly available dataset.

Models: -

A CNN (Convolutional Neural Network) was investigated alongside VGG16 - a pre-trained deep learning architecture, and the Random Forest Classifier - a traditional machine learning model.

Convolution Neural Network: A CNN algorithm requires less pre-processing compared to other traditional algorithms. CNN takes an input image and assigns importance (weights) to different features in the images for differentiability with other features in the image. These weights are assigned to each neuron that picks the specific feature when the image pixels are fed through it.

CNN algorithms are typically made up of a series of convolution and pooling layers, followed by one or more fully connected layers. The final layer in this architecture is determined by the type of activation function used, such as SoftMax for multiclass classification and sigmoid for binary classification.

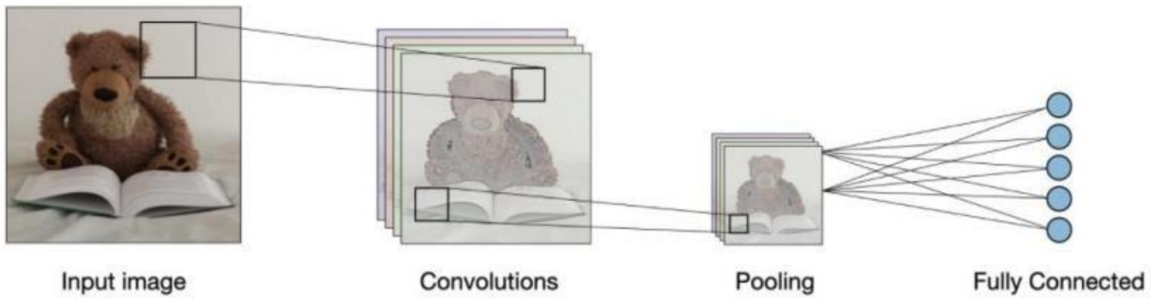


Figure1. Traditional CNN Layers

The convolution layer uses filters/kernel that performs convolution operation on the image as it scans through it. The hyperparameters that the convolution layers have are the number of filters and their size, stride, and padding. Stride is the value denoting the number of pixels by which the kernel will pass through. The output of the layer can either have the same size as the input or a reduced dimension and this is achieved by the padding parameter, followed by the pooling layer used for the discretization of the output data from the convolution layers. It is useful in reducing dimensionality and is also useful in extracting dominant features through spatial invariance. The hyperparameters are like the convolution layer. There are different types of pooling like max, average, global average, and global max and in the last stage, Fully connected layers are where each input is connected to all the neurons in the layer. It operates on a flattened input. They are usually found at the end of the CNN architecture where the flatten matrix goes through for the classification of images.

$$x_{ij}^{\ell} = \sum_{a=0}^{m-1} \sum_{b=0}^{m-1} \omega_{ab} y_{(i+a)(j+b)}^{\ell-1} .$$

An equation to calculate the Pre-Nonlinearity of the CNN Model

VGG16: VGG 16 was proposed for "The ImageNet Large Scale Visual Recognition Challenge (ILSVRC)," in which 14 million images must be classified into 1000 classes. The VGG is composed of 5 blocks of 2 or 3 convolution layers, followed by max-pooling layers, and 3 fully connected layers, followed by a SoftMax for classification.

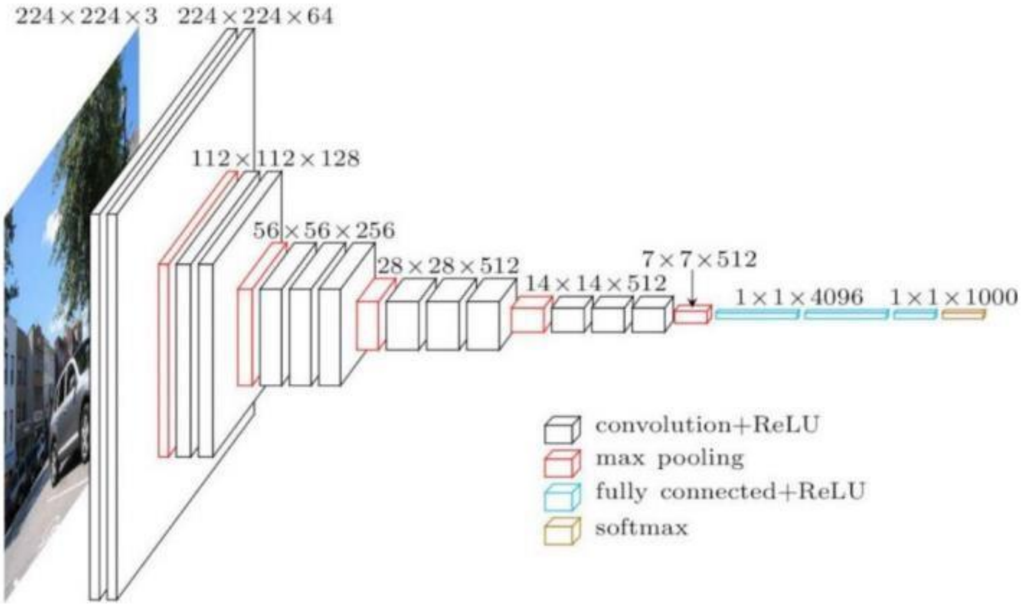


Figure 2. VGG16 Architecture

$$output = \frac{input - kernel\_size + 2 * padding}{stride} + 1$$

Random Forest Classifier: The random forest is made up of many individual decision trees that work together as an ensemble; each tree predicts a class, and the class with the most votes is the final prediction.

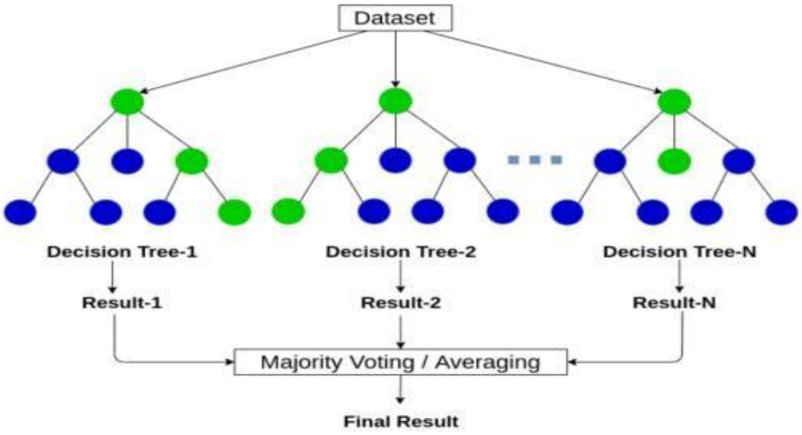
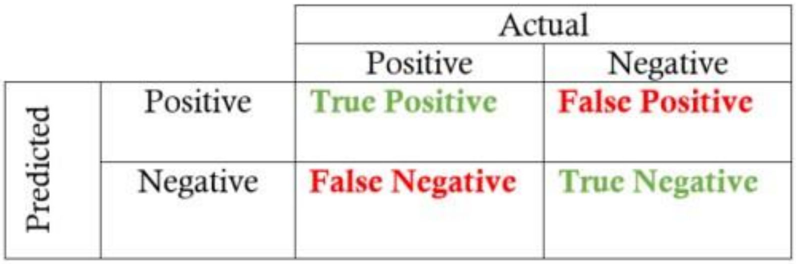


Figure 3. Random Forest Classifier Architecture

The hyperparameter used in the tuning of the random forest is:-

1. **n_estimator:** It represents the number of the decision trees to be used in the random forest algorithm.
2. **min_samples_split:** It is the minimum number of data points placed before the node is split.
3. **min_samples_leaf:** It is the minimum number of data points to be placed at each leaf node.
4. **max_features:** The maximum number of features considered while splitting a node.
5. **Criterion:** The function to measure the quality of a split. It is a tree-specific parameter.
6. **Bootstrap:** Methods used for sampling of data. If true, each tree has a different set of training data while false uses the same training data without replacement.

6. Performance Metrics: All models were tested on the test dataset after training. Accuracy, recall, and precision was used to validate the results. All the metrics are discussed further below.



		Actual	
		Positive	Negative
Predicted	Positive	True Positive	False Positive
	Negative	False Negative	True Negative

Figure 4. Confusion Matrix

In a binary classification of pneumonia and healthy chest X-rays, for a class, say pneumonia, True positive (TP) is the number of pneumonia images classified as pneumonia

True negative (TN) is the number of healthy images classified as healthy

False-positive (FP) is the number of normal images incorrectly identified as pneumonia, and

False-negative (FN) is the number of pneumonia images incorrectly classified as normal in a binary classification of pneumonia and healthy chest X-rays. These can also be calculated for multiclass classification.

- **Accuracy:** It shows how close are the predicted values to the known value.

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

- **Precision:** It shows how accurate the model is in terms of the predicted values.

$$Precision = \frac{TP}{(TP + FP)}$$

- **Recall:** It shows how accurate the model is in terms of the known values or it tells how many actual positives our model has captured through labeling them as positive.

$$Recall = \frac{TP}{(TP + FN)}$$

7. Methods:

Data Preprocessing: Each algorithm requires a different format of the input dataset; thus, each image must be preprocessed differently based on the model's requirements. The main parameters are the size and shape of the image that is used. Traditional CNN, pre-trained VGG16, and random forest models used RGB image formats with image sizes of 64x64, 224x224, and 128x128. The CNN and VGG16 input shapes were part of the format.

Input shape = [(number of images), (image height), (image width), (image channels)]

Each image was converted to a single vector and then appended together to form a matrix for a random forest. For training of the neural network, an adequate amount of the images is required, so the image dataset of 11695 images was split for training and validation of the model. The dataset was divided into 9:1 ratio, so we had 10525 images for training and 1170 for model validation.

Data Augmentation: The major problem is the lack of dataset images which is used for training and this lack of dataset, causes an overfitting problem for the model. To overcome this, we use a technique called Data Augmentation.

Data augmentation is a very powerful technique used to artificially create variations in existing data to expand an existing image data set, this creates new and different images from the existing image data set that represents a comprehensive set of possible images. This is done by applying different transformation techniques like zooming the existing image, rotating the existing image by a few degrees, shearing or cropping the existing set of images, etc.

This technique helps to increase the performance of the model by generalizing better images and thereby reducing overfitting.

Classification: The classification process was carried out with 5 architectures. Traditional CNN, VGG16, and random forest architectures were used. Then, with input from the feature extraction models, CNN and VGG16 were used as feature extraction models and Random Forest as classifiers.

The below figure depicts the traditional CNN architecture that was used. It is made up of three layers, which are followed by a fully-connected layer with SoftMax activation. The first layer is a convolution layer with 32 filters of 33 different sizes. The stride was set to 1 and the padding was the same. The rectified linear unit activation function was used (RELU). Following that, the output was batch normalized, which aids in the stabilization of the learning process and reduces the number of training epochs. After that, a max-pooling layer with a pool size of 22 and a stride of 2 was added. A single fully connected layer with three channels, the same as the class labels, was used. For training, the Stochastic Gradient Descent (SGD) optimizer was used.

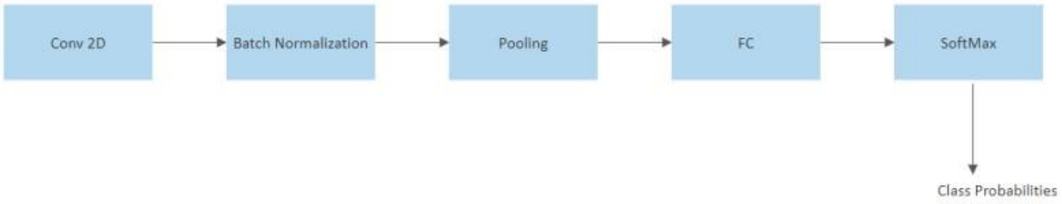
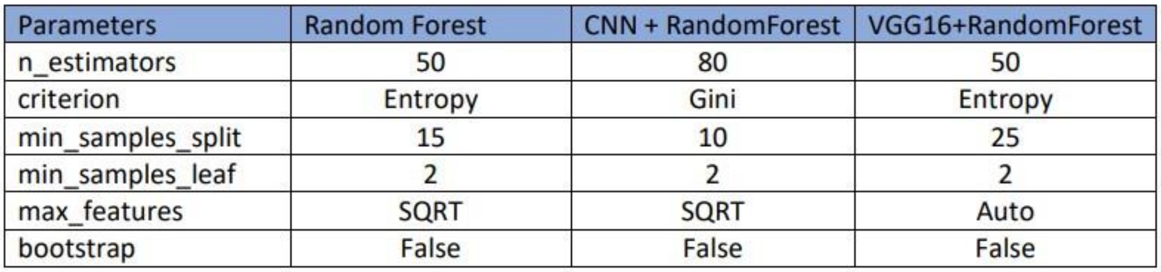


Figure 5. Schematic Representation of CNN model

The VGG16 architecture was used, with hyperparameter tuning in the fully connected layers. The standard VGG16 model includes 4096 channels in the first two fully connected layers and 1000 channels in the final layer, but in our project, we changed channels to 4096 and 2048 for this study. r. The activation function used was RELU for all the weighted layers and SoftMax in the last layer of fully connected layers. The optimizer used for VGG16 was the Adam optimizer.

The random forest classifier was used by optimizing the architecture's hyperparameters by implementing the maximization of the recall function. Later, neural networks without fully connected layers were used for feature extraction, and the random forest classifier was used instead of fully connected layers for image class label classification. The hyperparameter for each model was optimized for better results by maximizing the recall value. Table 1 shows the hyperparameters that were used.

Table 1. Hyperparameter for Random Forest

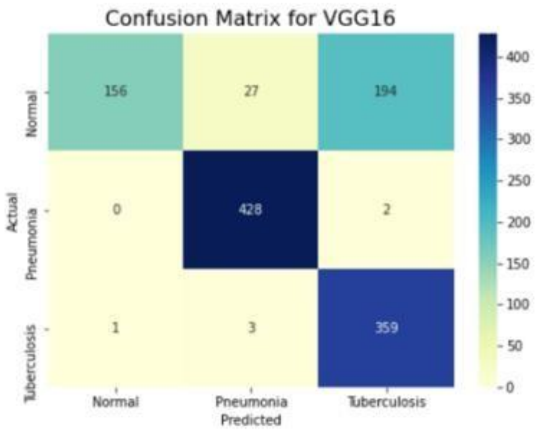


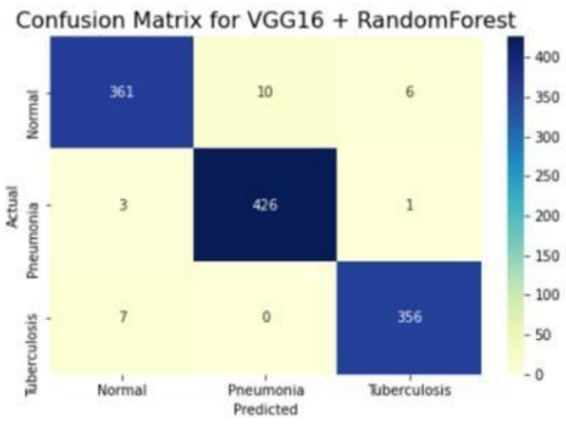
Parameters	Random Forest	CNN + RandomForest	VGG16+RandomForest
n_estimators	50	80	50
criterion	Entropy	Gini	Entropy
min_samples_split	15	10	25
min_samples_leaf	2	2	2
max_features	SQRT	SQRT	Auto
bootstrap	False	False	False

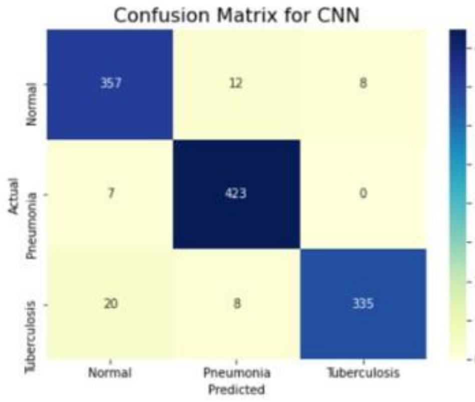
8. Results:

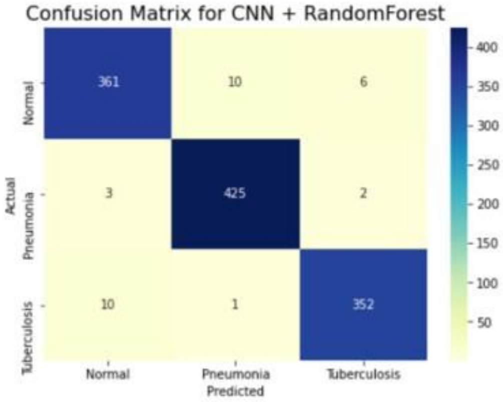
Performance Evaluation: Following the completion of the training, all the models were tested on the test data. The confusion matrix for all models was obtained to test the robustness of the models.

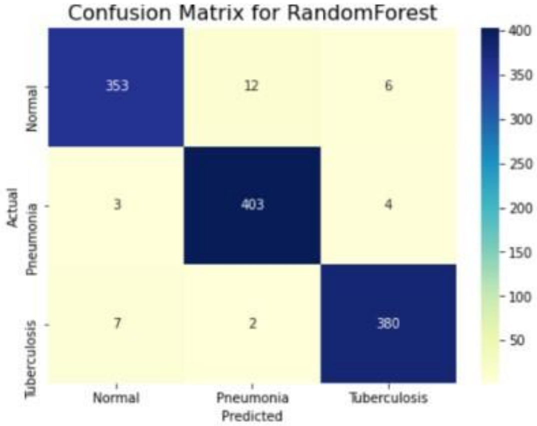
The confusion matrix was used to calculate the number of true positives, true negatives, false negatives, and false positives. The performance metrics were calculated using these values.











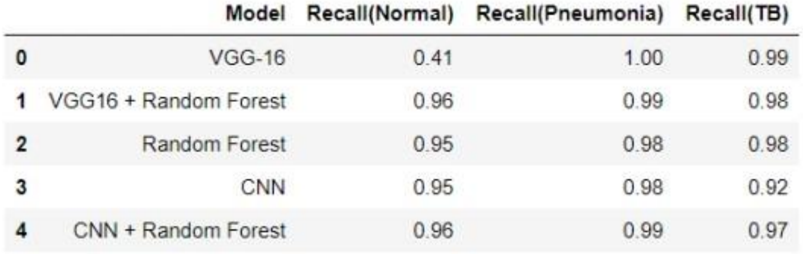
As we aim to reduce the false-negative number, it was discovered from the confusion matrices that the false-negative number for the class label "pneumonia" was very low when compared to the total number of pneumonia cases. A similar pattern was seen with the tuberculosis class label. However, these values had a high variance for the normal class label, as observed. In comparison to the total of 377 normal images, VGG16 had 221 false negatives. This value was observed to have reduced drastically for the other models in which they were lying very close to 20 false negatives.

The confusion matrix was used to calculate the error rate, which was found to be the highest (19.4 percent) for the VGG16 model and the lowest (2.31 percent) for the combined VGG16 and random forest model.

In a medical application, all patients who have the disease must be identified as sick (actual positive), but if they are not sick (predicted false), the cost of this can be extremely high if the disease is contagious. As a result, we seek to maximize model recall. From the above confusion matrices, we can see that the model VGG16 + Random Forest has outperformed all the other models overall. But looking at pneumonia and tuberculosis class, we see that VGG16 has outperformed others but has predicted most normal images as tuberculosis.

This is also demonstrated in Table 2, where the recall of normal images for the model VGG16 has a very low recall compared to other models.

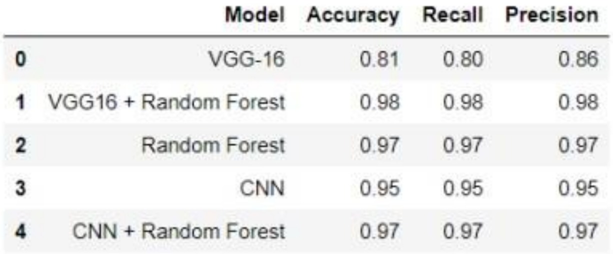
Table 2. Recall values



	Model	Recall(Normal)	Recall(Pneumonia)	Recall(TB)
0	VGG-16	0.41	1.00	0.99
1	VGG16 + Random Forest	0.96	0.99	0.98
2	Random Forest	0.95	0.98	0.98
3	CNN	0.95	0.98	0.92
4	CNN + Random Forest	0.96	0.99	0.97

Table 2 shows that for all models, the recall of normal images and TB is always less than the recall of pneumonia. The imbalance in the number of images in each class is to blame for this misclassification.

Table 3. Performance Metrics



	Model	Accuracy	Recall	Precision
0	VGG-16	0.81	0.80	0.86
1	VGG16 + Random Forest	0.98	0.98	0.98
2	Random Forest	0.97	0.97	0.97
3	CNN	0.95	0.95	0.95
4	CNN + Random Forest	0.97	0.97	0.97

9. Conclusion:

The VGG16 + Random Forest model is the best choice for clinical decision-making because of its hightest accuracy (98%) and overall recall (98%). The confusion matrix reveals that the false positive was greater than the false negative for each class, resulting in a small test error for class misclassification.

It is also worth noting that the traditional CNN and the pre-trained neural network VGG16 performed well for pneumonia and tuberculosis classes but not for the normal images. While the model with the random forest as a classifier and the neural network as a feature extractor outperforms the neural network classifier, Thus, a convolution neural network is used to extract the important features and details from a dataset.

10. Future Work:

Furthermore, because the visualization capabilities of these models have not been thoroughly investigated, we can investigate different visualization methods for these models, providing more insights into how the model works.

These methods may also aid in the tuning of these models' hyperparameters for a better result. In the future, it would be interesting to see approaches for more efficiently estimating the weights corresponding to different models, as well as a model that makes predictions while taking the patient's history into account.

11. Contributions:

Rahul Maddula-

1. Information Gathering.
2. Preprocessed the data for CNN and Random Forest.
3. Tested Algorithm.

Naga Surya Suresh-

1. Gathered information on the image dataset.
2. Preprocess the data for VGG16 and Random Forest.
3. Documented the project.

Pooja Pramod Kantrod-

1. Prepared and preprocessed the dataset.
2. Trained and validated algorithms.
3. Performed hyperparameter tuning.
4. Documented the project.

12. References:

[1] Hashmi, M. F., Katiyar, S., Keskar, A. G., Bokde, N. D., & Geem, Z. W. (2020), "Efficient Pneumonia Detection in Chest Xray Images Using Deep Transfer Learning. Diagnostics", MDPI Journal (Basel,
Switzerland), 10(6), 417

[2] Pasa, F., Golkov, V., Pfeiffer, F. et al, "Efficient Deep Network Architectures for Fast Chest X-Ray Tuberculosis Screening and Visualization", Sci Rep 9, 6268 (2019).

[3] Simonyan, K. & Zisserman, A. (2014), "Very Deep Convolutional Networks for Large-Scale Image Recognition", CoRR, abs/1409.1556.

[4] Liang G. & Zheng L, "A transfer learning method with deep residual network for pediatric pneumonia diagnosis. Computer Methods and Programs in Biomedicine", 187 pp. 104964 (2020) pmid:31262537

[5] Kermany D., Zhang K. & Goldbaum M. Labeled, "Optical Coherence Tomography (OCT) and Chest X-Ray Images for Classification", (Mendeley,2018).

[6] Albahli S., Rauf H., Algosaibi A. & Balas V. "AI-driven deep CNN approach for multi-label pathology classification using chest X-Rays", PeerJ Computer Science. 7 pp. e495 (2021) pmid:33977135.

[7] Rahman T., Chowdhury M., Khandakar A., Islam K., Islam K., Mahbub Z., et al, “Transfer learning with deep convolutional neural network (CNN) for pneumonia detection using chest X-ray”, Applied Sciences. 10, 3233 (2020).

[8] Zubair S, “An Efficient Method to Predict Pneumonia from Chest X-Rays Using Deep Learning Approach”, The Importance of Health Informatics In Public Health During A Pandemic. 272 pp. 457 (2020)

[9] Rajpurkar P., Irvin J., Zhu K., Yang B., Mehta H., Duan T., et al. & Others Chexnet, “Radiologistlevel pneumonia detection on chest x-rays with deep learning”, ArXiv Preprint ArXiv:1711.05225. (2017)

[10] Albahli S., Rauf H., Arif M., Nafis M. & Algosaibi , “An Identification of thoracic diseases by exploiting deep neural networks”, Neural Networks. 5 pp. 6 (2021)

[11] Sharma H., Jain J., Bansal P. & Gupta S, “ Feature extraction and classification of chest x-ray images using cnn to detect pneumonia”, 2020 10th International Conference On Cloud Computing, Data Science & Engineering (Confluence). pp. 227-231 (2020)

[12] Stephen O., Sain M., Maduh U. & Jeong D, “An efficient deep learning approach to pneumonia classification in healthcare”, Journal Of Healthcare Engineering. 2019 (2019) pmid:31049186

[13] Kundu R., Basak H., Singh P., Ahmadian A., Ferrara M. & Sarkar R, “ Fuzzy rank-based fusion of CNN models using Gompertz function for screening COVID-19 CT-scans”, Scientific Reports. 11, 14133 (2021,7), pmid:34238992

[14] Mahmud T., Rahman M. & Fattah S. CovXNet: , “A multi-dilation convolutional neural network for automatic COVID-19 and other pneumonia detection from chest X-ray images with transferable multireceptive feature optimization”,Computers In Biology And Medicine. 122 pp. 103869 (2020) pmid:32658740

[15] Liu N., Wan L., Zhang Y., Zhou T., Huo H., Fang T, “Exploiting convolutional neural networks with deeply local description for remote sensing image classification”, IEEE Access. 2018;6:11215–11228. doi: 10.1109/ACCESS.2018.2798799.

[16] Abiyev R.H., Ma’aitah M.K.S, “Deep convolutional neural networks for chest diseases detection”,J. Healthc. Eng. 2018; 2018:4168538. doi: 10.1155/2018/4168538.

[17] Rajaraman S., Candemir S., Kim I., Thoma G., Antani S,”Visualization and interpretation of convolutional neural network predictions in detecting pneumonia in pediatric chest radiographs”, Appl. Sci. 2018;8:1715. doi: 10.3390/app8101715.

[18] Lakhani P., Sundaram B, “Deep learning at chest radiography: Automated classification of pulmonary tuberculosis by using convolutional neural networks”, Radiology.2017;284:574–582.doi: 10.1148/radiol.2017162326.

[19] Toğaçar M., Ergen B., Cömert Z, “A deep feature learning model for pneumonia detection applying a combination of mRMR feature selection and machine learning models”, IRBM.2019doi: 10.1016/j.irbm.2019.10.006.

[20] Ayan E., Ünver H.M, “Diagnosis of Pneumonia from Chest X-Ray Images Using Deep Learning”, Proceedings of the 2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT); Istanbul, Turkey. 2–26 April 2019; pp. 1–5.

[21] Ankita Shelke, Madhura Inamdar, “Chest X-ray Classification Using Deep Learning for Automated COVID-19 Screening**”,** SN Computer Sci. 2021; 2(4): 300. Published online 2021 May 26. doi: 10.1007/s42979-021-00695-5