International Journal of Interdisciplinary Innovative Research &Development (IJIIRD)
ISSN: 2456-236X
Vol. **05** Issue **02** | **2021**

Lung Disease Classification using Chest X-rays

[1]Naveen Gummala, [2]S. Prasad Babu Vagolu, [3]Srinivasa Rao B
[1, 2, 3]Gitam Deemed to be University, Visakhapatnam

ABSTRACT

Lung Disease is one of the most common maladies around the world. An accurate lung cancer classifier could speed up and reduce costs of lung cancer screening, allowing for more widespread early detection and improved survival. In this report, deep learning architectures will be applied on the dataset from Lung Nodule Analysis 2016 which consists of 3D CT scan of patients. The initial task is to use segmentation techniques to detect nodule images from non-nodule images. Data Preprocessing and Data Augmentation constitute two most important steps in the Image preprocessing. Our project aims to compare performance of various low memory, light- weight deep neural net (DNN) architectures for biomedical image analysis. This will involve using networks including DenseNet121, MobileNetV2, EfficientNetB0, InceptionResNetV2,2D SqueezeNet, 2D MobileNet for binary classification in order to detect the presence of lung cancer in patient CT scans of lungs with and without earlystage lung cancer.

1. INTRODUCTION

The world is changing so fast that the pressure on health is increasing, the bad changes of climate, the environment, the life way of human, ... also increase the risk as well as diseases for people. One of the issues that we will focus on in this article is lung diseases. About 3.2 million people succumbed in 2015 to chronic obstructive pulmonary disease (COPD), caused mainly by smoking and pollution, while 400,000 people died from asthma can get, here is just one example of diseases we can save if we find them out earlier. With the technology machine and computer power, the earlier identification of diseases, particularly lung disease, we can be helped to detect earlier and more accurately, which can save many and many people as well as reduce the pressure on the system. The health system has not developed in time with the development of the population. With the power of computers as well as the large amount of data being released to the public, this is a good time to contribute to solving this problem. Wishing to contribute more to the community, helping those who are not able to pay for medical expenses, I hope that my solution can contribute to reducing medical costs, the development of computer science for medical projects.

2. LITERATURE SURVEY

For the detection of pulmonary diseases chest radiography is always required to identify pulmonary problems. Diseases such as tuberculosis, pneumonia and lung cancer are a major threat to human health. Thus, in (Khobragade et al., 2016) proposes pulmonary segmentation, extraction of characteristics and its classification using an artificial neural network for the detection of pulmonary diseases. (Khobragade et al., 2016) a simple image processing technique with intensity-based method was used, and a method based on the discontinuity to detect pulmonary limits, in this way, statistical and geometric characteristics were extracted. Neural networks were used feed forward and back propagation to detect diseases.

Pneumonia is one of the leading causes of infant mortality. In developing countries there is little infrastructure and doctors in rural areas to provide the necessary diagnosis. Therefore, in (Barrientos et al., 2016) proposes a method for automatic diagnosis using ultrasonography of the lungs. The approach presented is based on the analysis of patterns in rectangular segments in the image of the ultrasonography. The specific characteristics and characteristic vectors were obtained and classified by a standard neural network. In (Barrientos et al., 2016) I obtained a sensitivity of 91.5 % and specificity of 100 % but were extracted from a single patient and only included in the test or in the training set.

Many researchers have developed several algorithms for the diagnosis of lung diseases through sound. One of the parameters used for the detection of pulmonary sound is entropy, so there are differences in the sound of a normal respiratory system and a system with pathologies. In the article (Rizal et al., 2017) discurses several measures of entropy for a classification of pulmonary sounds. The result in (Rizal et al., 2017) shows that the use of a single entropy could not achieve high accuracy, so 7 entropies were used and guaranteed 94.95 % accuracy using multilayer perceptron.

In paper (Rodrigues et al., 2018) suggests a Structural Co-occurrence Matrix (SCM) approach to classify malignant nodules or benign nodules and also their level of malignancy. The structural cooccurrence matrix

technique was applied to extract characteristics of the nodule images and classify them. The SCM was applied in gray scale and images of the Hounsfield unit with four filters, creating eight different configurations. The classification stage used classifiers known as the multilayer perceptron, support vector machine, k-Nearest Neighbors algorithm and were applied in two tasks: (i) to classify the nodule images as malignant or benign, (ii) to classify the nodules pulmonary lesions at the level of malignancy (1 to 5). O (Rodrigues et al., 2018) had a result of 96.7 % for precision and F-score measurements in the first task and 74.5 % accuracy and 53.2 % F-score in the second task. The (Santosh and Antani, 2018) has proposed an idea that takes into account the alterations of the right and left lung region in terms of symmetry and automated the chest X-ray system for the evidence of tuberculosis. The proposed method is the observation of radiological exams leading to bilateral comparisons in the lung field.

In (Santosh and Antani, 2018) analyzed the symmetric lung region using a multiplescale shape feature, as well as border texture characteristics. Three different types of classifiers were used: Bayesian network, multilayer network percepClassification of Images of Childhood Pneumonia using Convolutional Neural Networks 113 tron and random forest. The results obtained with an abnormality detection accuracy of 91 % and area under the ROC curve of 0.96. Many researchers use various methods for detecting diseases based on lung sound, for example the use of entropy measurement. Sound of pulmonary snoring is a sound that is discontinuous, of short duration and appears in the inspiratory, expiratory or in both cases. Thus, in (Rizal et al., 2016) the Tsallis entropy was used as the characteristic extraction method to classify lung sounds. The results were achieved using at least three Tsallis entropy values with q = 2, 3, and 4 with MLP as a classifier and three-fold cross validation at an accuracy of 95.35 %, sensitivity of 90.48 % and 100 % specificity, were achieved. Lung cancer accounts for 26 % of all cancer deaths in 2017, accounting for more than 1.5 million deaths. Thus, in (d. Nobrega et al., 2018) is ´ proposed to explore the performance of deep transfer learning to classify pulmonary nodule malignancies.

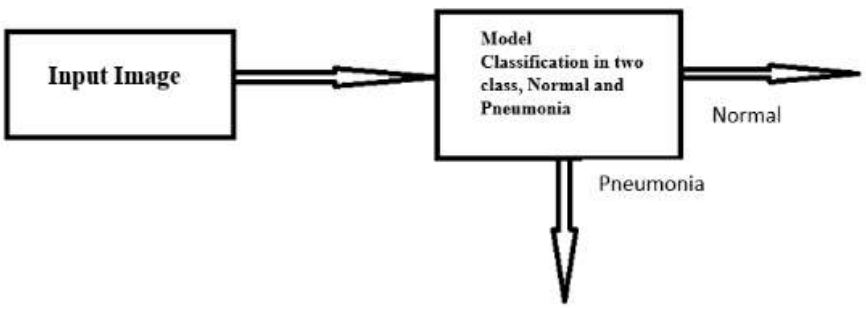
In this work, we have used a convolutional neural network such as VGG16, VGG19, MobileNet, Xception, InceptionV3, ResNet50, Inception-ResNet-V2, DenseNet169, DenseNet201, NASNetMobile and NASNetLarge, where they were used to extract parameters from an image database of the lung. The characteristics were classified using Naive Bayes, Perceptron multilayer, support vector machine, Near Neighbors and Random Forest. The results obtained in (d. Nobrega et al., 2018) were 88.41 % of ACC and ´ 93.19 % of AUC. (Paing and Choomchuay, 2017) has the objective of detecting pulmonary nodules from a series of digitized CT images. The threshold and morphological operations of Otsu are applied for the segmentation of nodules.

In (Paing and Choomchuay, 2017), after segmentation, objects that can not be nodes are discarded. In view of this, multilayer Perceptron was used for the classification and 95 % accuracy was achieved. In the work (Kermany et al., 2018) a method was proposed using Convolutional Neural Networks with a transfer learning technique. Transfer learning has proven to be a highly effective technique, particularly when confronted in domains with limited data.

3. PROBLEM IDENTIFICATION AND OBJECTIVE

Recently a large dataset of X-ray lung data was public on Kaggle followed by labeled lung disease data. This is a good condition for me to implement this project. In this project I will conduct a study and analysis of this data set, then apply Machine Learning and Deep Learning to predict that the patient has a lung disease. This project is a binary classification with input is patient's data (age, gender, X-ray images, View Position) and output is found diseases or not.

The difficulty is a new dataset, and I will be one of the pioneers to learn it, my analysis is that thisis a large dataset but has never been processed full, data has a lot of noise, and X-ray of the lung is not likely to provide enough information to assess whether a patient may be ill. I will use Machine Learning as well as Deep Learning to process data as well as create models for diagnosing patients. My keys point here will be: combining the processing of patient information with data from X-rays, using Keras and tensorflow with the well-known pre-trained models, DenseNet121, MobileNetV2, EfficientNetB0, InceptionResNetV2,2D SqueezeNet, 2D MobileNet.



4. SYSTEM METHODOLOGY

4.1 Dataset Description

Dataset is collected from Kaggle(https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia). In the dataset used for training and validation, it contained 5863 X-ray (JPEG) images and two categories: Normal and Pneumonia. A total of 5863 patients (Pneumonia and normal) The radiographic images were from pediatric patients one to five years old from the Medical Center in Guangzhou. In this way, radiographs were performed as part of clinical care. All images in dataset underwent a treatment in order to remove all low-quality scans, as well as being classified by two specialist physicians and by a third-party specialist, in order to prevent any misclassification.

4.1.1 Training Dataset

The sample of data used to fit the model. The actual dataset that we use to train the model (weights and biases in the case of a Neural Network). The model *sees* and *learns* from this data.

4.1.2 Validation Dataset

The sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyperparameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration.

The validation set is used to evaluate a given model, but this is for frequent evaluation. We, as machine learning engineers, use this data to fine-tune the model hyperparameters. Hence the model occasionally *sees* this data, but never does it "*Learn*" from this. We use the validation set results, and update higher level hyperparameters. So the validation set affects a model, but only indirectly. The validation set is also known as the Dev set or the Development set. This makes sense since this dataset helps during the "development" stage of the model.

4.1.3 Test Dataset

The sample of data used to provide an unbiased evaluation of a final model fit on the training dataset. The Test dataset provides the gold standard used to evaluate the model. It is only used once a model is completely trained (using the train and validation sets). The test set is generally what is used to evaluate competing models (For example on many Kaggle competitions, the validation set is released initially along with the training set and the actual test set is only released when the competition is about to close, and it is the result of the the model on the Test set that decides the winner). Many a times the validation set is used as the test set, but it is not good practice. The test set is generally well curated. It contains carefully sampled data that spans the various classes that the model would face, when used in the real world.

- Number of pictures in the training set: 5232
- Number of pictures in the test set: 624
- Number of pictures in the validation set: 624

4.2 Pneumonia (Bacterial and Viral) and Normal

For the diagnosis of pneumonia, the alveoli become filled with secretion and appear as a white spot on the chest radiograph. Pulmonary consolidation means that the pulmonary alveoli are filled with inflammatory fluid. In radiography, pulmonary consolidation corresponds to an opacity, that is, the whitish area.

(I) Normal (II) Pneumonia

The dataset is organized into 3 folders (train, test, val) and contains subfolders for each image category (Pneumonia/Normal). There are 5,863 X-Ray images (JPEG) and 2 categories (Pneumonia/Normal). Chest X-ray images (anterior-posterior) were selected from retrospective cohorts of pediatric patients of one to five years old from Guangzhou Women and Children's Medical Center, Guangzhou. All chest X-ray imaging was performed as

part of patients' routine clinical care. For the analysis of chest x-ray images, all chest radiographs were initially screened for quality control by removing all low quality or unreadable scans. The diagnoses for the images were then graded by two expert physicians before being cleared for training the AI system. In order to account for any grading errors, the evaluation set was also checked by a third expert.
Here we import data and use 27 pretrained models from Keras and follow the steps below:

- Data pre-processing and visualization
- Load the Images with a generator and Data Augmentation
- Test 27 canned architectures with pre-trained weights
- Train the best architecture
- Examples of prediction

4.3 The pretrained Keras modules used here are

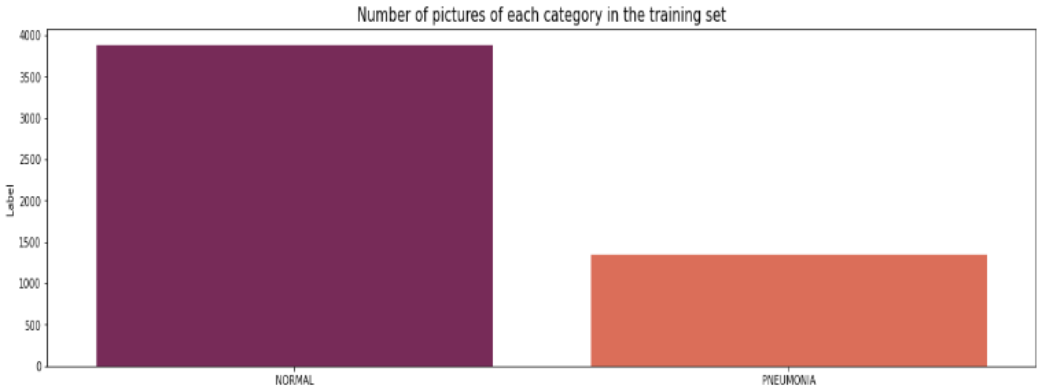
Keras Applications are canned architectures with pre-trained weights.

4.3.1 Modules

densenet module: DenseNet models for Keras.
efficientnet module: EfficientNet models for Keras.
imagenet_utils module: Utilities for ImageNet data preprocessing & prediction decoding.
inception_resnet_v2 module: Inception-ResNet V2 model for Keras.
inception_v3 module: Inception V3 model for Keras.
mobilenet module: MobileNet v1 models for Keras.
mobilenet_v2 module: MobileNet v2 models for Keras.
mobilenet_v3 module: MobileNet v3 models for Keras.
nasnet module: NASNet-A models for Keras.
resnet module: ResNet models for Keras.
resnet50 module: Public API for tf.keras.applications.resnet50 namespace.
resnet_v2 module: ResNet v2 models for Keras.
vgg16 module: VGG16 model for Keras.
vgg19 module: VGG19 model for Keras.
xception module: Xception V1 model for Keras.

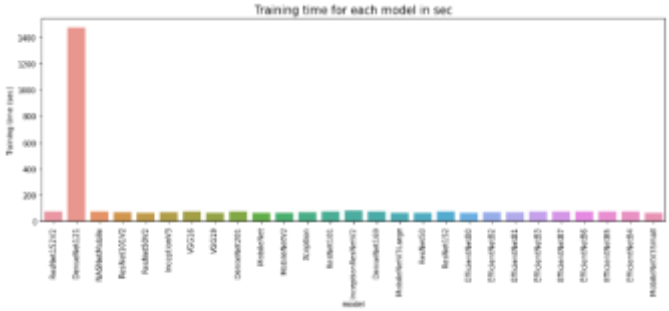
	Filepath	***Label***
0	*/content/drive/MyDrive/chest_xray/train/PNEUMO...*	*PNEUMONIA*
1	*/content/drive/MyDrive/chest_xray/train/PNEUMO...*	*PNEUMONIA*
2	*/content/drive/MyDrive/chest_xray/train/PNEUMO...*	*PNEUMONIA*
3	*/content/drive/MyDrive/chest_xray/train/PNEUMO...*	*PNEUMONIA*
4	*/content/drive/MyDrive/chest_xray/train/NORMAL...*	*NORMAL*
5	*/content/drive/MyDrive/chest_xray/train/NORMAL...*	*NORMAL*

4.4 Display the number of pictures of each category in the training set



4.5 Accuracy on Test set

4.6 Training time for each model in sec

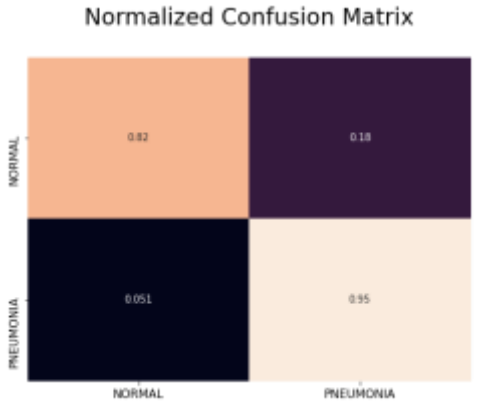


Best model: ResNet152V2
Accuracy on the test set: 90.22%

5. RESULTS AND METRICS

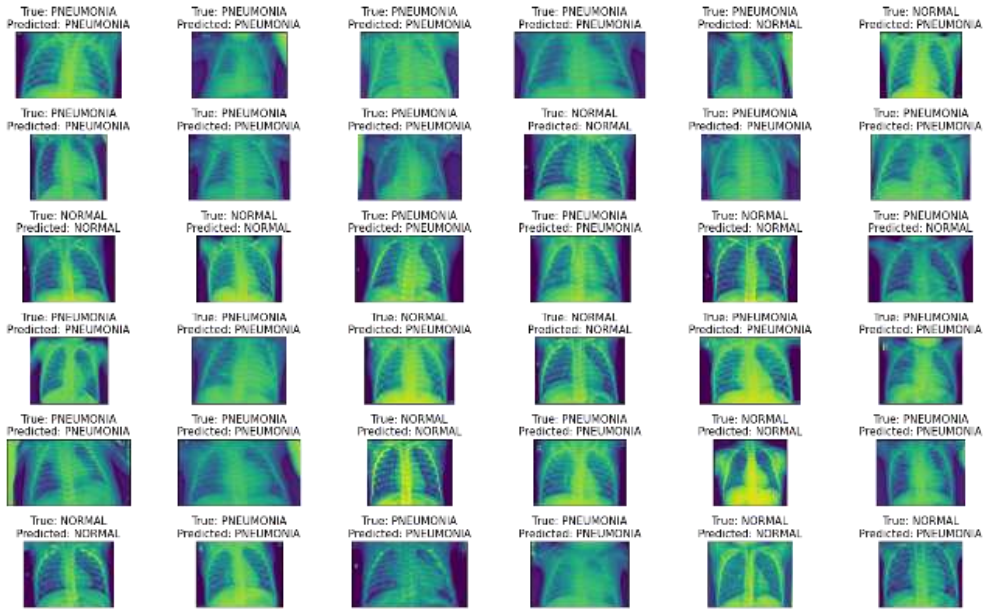
As visible on the classification report, the recall for the label "pneumonia" is very high with 97%. The accuracy of the prediction of "normal" isn't very high with around 68%. Explained with other words in a simplified way: Most of the people with pneumonia are identified and some people without pneumonia are wrongly identified as having it. In the medicinal field, this is normal, because it is better to create models, which wrongly identity a sickness but don't miss people who are sick for real.

	precision	**recall**	**f1-score**	**support**
NORMAL	0.91	0.82	0.86	234
PNEUMONIA	0.90	0.95	0.92	390
accuracy			0.90	624
macro avg	0.90	0.89	0.89	624
weighted avg	0.90	0.90	0.90	624



The confusion matrix is an array that contains correct and incorrect predictions of the algorithm and the actual situation. As shown in table

• True Positive: Number of people who actually have pneumonia according to the algorithm.
• False Negative: Number of people who are actually with pneumonia but categorized as healthy according to the algorithm



6. CONCLUSION

In this project we have demonstrated a comparison with the work of in the detection and classification of images for the detection of pneumonia from the chest X-ray of patients. The Convolutional Neural Networks was used to train the neural network and, for the validation of the model, Cross validation was used. The classification model presented was efficient in the classification, obtaining an average accuracy of 95.30 % in the tests against 92.8 % of the work.

We aim to extend the project further to adapt these deep learning architectures to predict using 3D volumetric slices. Moreover, the bright pixels usually corresponded with the location of cancerous nodules, so it could be possible to extend our current model to determine the exact location of the cancerous nodules. A part of the future work will also involve experimentation with various segmentation processes (K- means, Watershed) to improvise the 2D models.

7. REFERENCES

[1] https://medicalxpress.com/news/2017-08-lung-diseases-million.html
[2] Andrew Ward, Nicholas Bambos. Quantum Annealing Assisted Deep Learning for Lung Cancer Detection. http://cs231n.stanford.edu/reports/2017/pdfs/534.pdf
[3] Albert Chon, Niranjan Balachandar. Deep Convolutional Neural Networks for Lung Cancer Detection. http://cs231n.stanford.edu/reports/2017/pdfs/518.pdf
[4] Kingsley Kuan, Mathieu Ravaut, Gaurav Manek, Huiling Chen, Jie Lin, Babar Nazir, Cen Chen, Tse Chiang Howe, Zeng Zeng, Vijay Chandrasekhar. Deep Learning for Lung Cancer Detection: Tackling the Kaggle Data Science Bowl 2017Challenge.
[5] https://arxiv.org/abs/1705.09435
[6] https://www.nature.com/articles/srep46479
[7] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5569872/
[8] https://stephenson.pure.elsevier.com/en/publications/computer-aided-lung-cancer-diagnosis-with-deep-learning-algorithm
[9] https://www.researchgate.net/publication/301651102_Computer_aided_lung_cancer_diagnosis_ with_deep _learning_algorithms
[10] Matrix capsules with EM routing. https://openreview.net/forum? id=HJWLfGWRb&noteId=HJWLfGWRb
[11] Sara Sabour, Nicholas Frosst, Geoffrey E Hinton. Dynamic Routing Between Capsules.
[12] https://arxiv.org/abs/1710.09829
[13] NIH sample Chest X-rays dataset, https://www.kaggle.com/nih-chest-xrays/sample
[14] NIH full Chest X-rays dataset, https://www.kaggle.com/nih-chest-xrays/data
[15] Max Jaderberg, Karen Simonyan, Andrew Zisserman, Koray Kavukcuoglu. Spatial Transformer Networks. https://arxiv.org/abs/1506.02025
[16] Traffic Sign Recognition using CNN with Learned Color and Spatial Transformation. https://github.com/ hello 2all/ GTSRB_Keras_STN/blob/master/capstone_report.md
[17] CapsNet-Keras. https://github.com/XifengGuo/CapsNet-Keras