ISSN 2319 – 7595
Volume 8, No.2, March - April 2019
International Journal of Information Systems and Computer Sciences
Available Online at http://warse.org/IJISCS/static/pdf/file/ijiscs12822019.pdf
https://doi.org/10.30534/ijiscs/2019/12822019

XRAY AI: Lung Disease Prediction Using Machine Learning

Justin Monsi[1], Justine Saji[2], Keerthy Vinod[3], Liya Joy[4] , Jis Joe Mathew[5]
[1]Amal Jyothi College of Engineering, Kottayam, India, justinmonsi@cs.ajce.in
[2]Amal Jyothi College of Engineering, Kottayam, India, justinesaji@cs.ajce.in
[3]Amal Jyothi College of Engineering, Kottayam, India, keerthyvinod@cs.ajce.in
[4]Amal Jyothi College of Engineering, Kottayam, India, liyajoy@cs.ajce.in
[5]Amal Jyothi College of Engineering, Kottayam, India, jisjoemathew@amaljyothi.ac.in

ABSTRACT

The chest X-ray is a quick and effective test that has been useful for decades to help doctors to view vital organs. When focused on the chest, it can help spot abnormalities or diseases f the airways, bones, heart and lungs. The chest X-ray dataset currently available consists of details regarding 14 diseases. The lack in publically available datasets creates a difficulty in providing more Computer Aided Detection(CAD) in real world medical science with chest X-rays.
In this paper we present the NIH chest X-ray dataset comprised of 112,120X-ray images with disease labels from 30,000 unique patients. The labels are expected to be more than 90% accurate and we perform weakly supervised learning with this dataset. The deep convolutional neural network is used to recognize and locate the common disease patterns and we try to integrate the automated detection system with normal hospital management system.

Key words: Artificial Intelligence, Deep Learning, Pulmonary Diseases, Computer Aided Detection

1. INTRODUCTION

Deep learning is a type of machine learning where we learn from large amount of data which involves artificial intelligence, neural network and algorithm inspired by the human brain. It is the new perimeter for enterprise application and is a great assurance for development in almost every field. We combine this knowledge and apply machine learning as a real-world tool deployed to solve the problems of medical big data, having wide complexity and difficulty in handling. Identifying and diagnosing diseases are one of the many healthcare challenges we apply this knowledge to [8].

1.1 ResNet

ResNet is short for Residual Network which is used to facilitate the training of networks that are deep. It is more speed/memory efficient than dense network. The network we work with is ResNet 50 [5] [6].
The data we discuss are anonymized chest X-ray images and their corresponding data. The first challenge is to make sense of data using machine learning. This requires vigilant observation and proper comprehension of pathology. The basic idea is to teach the system to peruse and study to potentially discover common thoracic diseases.

1.2 Goals

- to create a system with minimum error that the doctors can use to get a second opinion.
- To reduce misdiagnosis and thus reducing fatal consequences.

2. RELATED WORKS

2.1 Computer Aided Screening of Pulmonary Diseases.

Focused on pulmonary Tuberculosis the US National Library of Medicine has release two datasets of postero-anterior (PA) chest radiographs [4][7]. It included normal X-rays as well as Tuberculosis infected X-rays and their associated readings. This dataset has been used for classification, experiment and lung siltation. CAD system creates a categorization of likely and unlikely cases of Tuberculosis. The performance of the system is in an automated outlook and the shape and texture of the images are evaluated to produce a classifier.

2.2 Chest X-ray Analysis Of Lung Cancer.

DenseNet 121 by G.Huang et.al. combined with transfer learning scheme was used to assist in screening and further diagnosing lung cancer using chest X-ray images. It bears above 70% accuracy [3]. The JSRT(Japanese Society of Radiological Technology) dataset containing images, both serious and benign cases,are used.

Due to the lack of availability of a large dataset the model was first trained on a lung nodule dataset and then a lug cancer dataset. Transfer learning is applied several times to improve the performance. The image data needs to be normalized, resized and filtered before this. There is a base model and further retrained base models. In the initial classification images with and without nodule are separated and further on we classify benign and damaging cases. The strategy used in this system creates the model with higher mean accuracy and mean sensitivity but low standard deviation. Further works can be done on by adding on more features like family history, rate of smoking for more accuracy.

3. METHODOLOGY

Here we are discussing the methodology used in our study. The dataset used is described followed by the data preparation process. Accordingly, the sanitization process and the model is explained.

A.Dataset

NIH X-ray Dataset: The NIH Clinical Center released over one million anonymized chest X-ray images and their associated data of more than 30,000 patients [1]. It contains 112,120 frontal chest X-ray images that includes 14 different thoracic diseases. The label of these 14 diseases are Atelectasis, Cardiomegaly, Consolidation,Edema, Effusion, Emphysema, Fibrosis, Hernia, Infiltration, Mass, Nodule, Pleural Thickening, Pneumonia and Pneumothorax. A statistics of patient gender versus the 14 diseases is

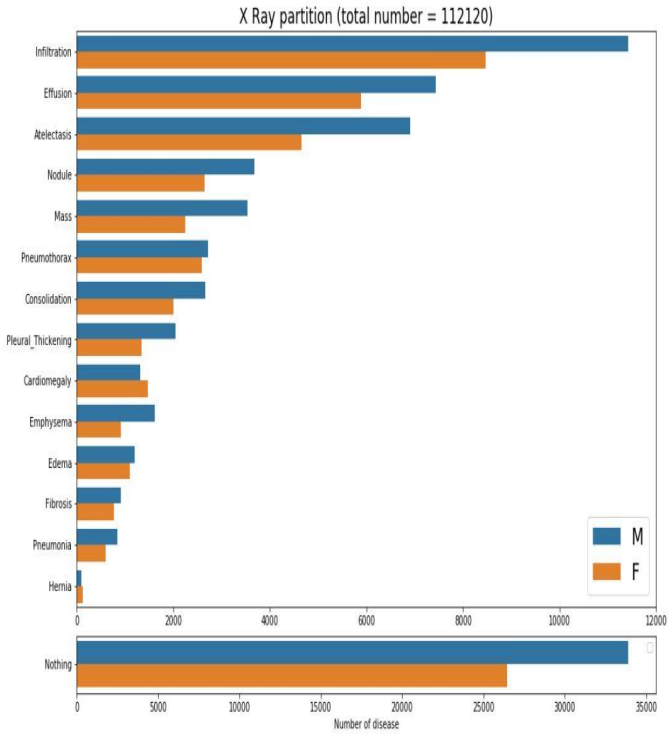

Figure 1. Dataset Overview: Patient gender analytics for 14 diseases

shown in Figure 1**.** The images are of size 1024 x 1024 pixels. We have to apply knowledge of pathology and
anatomy for further observation on this dataset to create a consistent technique[2].

Table 1.Training-Validation-Test Set

ChestX-ray	Training	Validation	Test
Atelectasis	2190	468	468
Cardiomegaly	490	105	104
Consolidation	693	149	148
Edema	356	76	76
Effusion	2115	453	453
Emphysema	471	101	101
Fibrosis	287	61	61
Infiltration	5277	1131	1131
Mass	1202	258	257
Nodule	1399	300	300
Pleural Thickening	562	120	120
Pneumonia	169	36	37
Pneumothorax	1170	250	251
No Findings	30688	6576	6576

B.Data Sanitization

We classify the images from the whole dataset according to different labels of these 14 diseases. There is also a label for cases of people having no identified diseases. Data is further randomly split into training, validation and test set, as shown in Table 1. Here we only use 67,346 of the 112,120 available images. There is no overlap of cases that is, multiple pathologies are ignored in our research.
After sanitization, case involving Hernia is ignored for further easiness of the system as the data available for it is scarce.

C.Data Preparation

The steps for data preparation is as follows:

- Step 1: resize the images to parallel the input of the model. Images are changed from 1024 x 1024 to 224 x 224 pixels.
- Step 2: the image color is normalized by converting it to grey scale.

An example of image after preparation is shown below in Figure 2.

D.Architecture of the model and training

Out of all available convolutional neural network, the model that to be used is ResNet 50. It is a residual network build to solve the problem of degradation. We pileup many residual blocks to form the residual network. There are 50 layers having 3 x 3 convolutions for mapping higher dimensions and 1 x 1 for lower dimensions. This CNN works efficiently on image

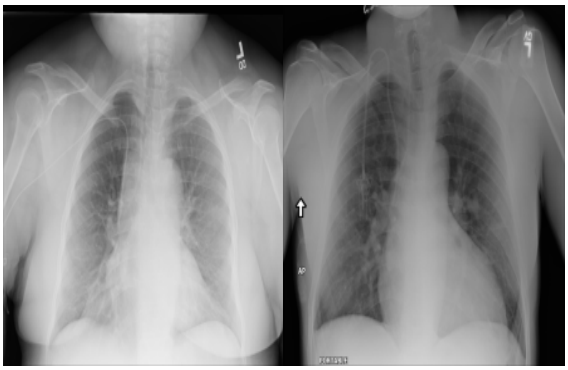

Figure 2. Illustration of data preparation

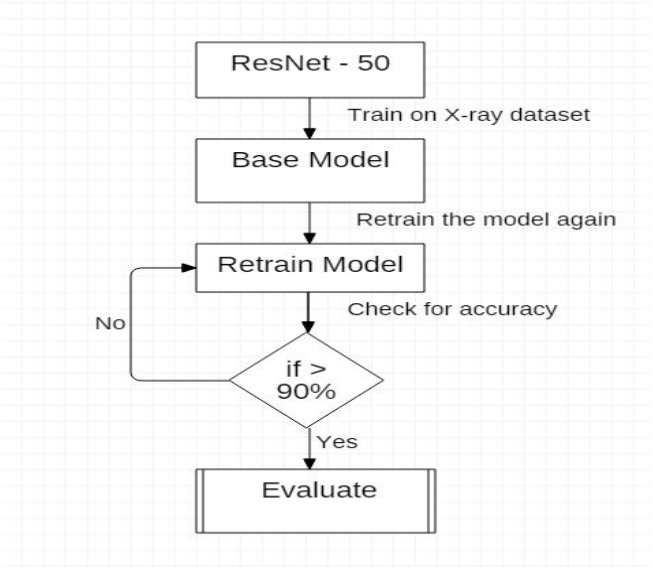

Figure 3. The training process

datasets[9]. It processes an input image having size 224 x 224 pixels. The initial training process takes a minimum of 70 hours. Through the training process we make the model more specific to our motive. The learning involves extracting features from the image and this process replicates with multiple epochs[10]. With each epoch atleast 40 of the images must be processed. This makes the system intelligent to predict a disease out of the available labels. The process is described in Figure 3. Transfer learning technique where one learns from stored data while applying it on another at the same time is used.
The system should not become too intelligent as it will lead to over fitting. The language used for creating the model must be chosen such that it will facilitate the process. Usually libraries in Python are used. Once integrated with the hospital management system, the doctor can test the image using a platform and the automated system will generate a tag that matches the most with the given input.

4. RESULTS AND CONCLUSION

We evaluate the performance of the system based on accuracy. The degree to which one can identify which cases are positive and which are negative is what we mean by accuracy. The model is said to have low error when it has high accuracy. It should be noted that the model should not overfit. After

completing multiple iteration on the model, the accuracy gets better with more and more data.

In our paper, we explored a system for detecting lung diseases from NIH X-ray dataset. We filtered the data, prepared into input to the network and train the model several times making it learn and finally predict the disease. In a hospital management environment this can be utilized as a mean for verification for the doctors.

Hereafter, we can add on features to this including lung cancer detection and other radiology image detection. The accuracy of the system can be enhanced by attaching attributes like rate of smoking, family history etc to the datalist entry.

REFERENCES

[1] **National Institutes of Health Chest X-Ray Dataset,**2018[Online]Availablle: https://www.kaggle.com/nih-chest-xrays

[2] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, Ronald M. Summers, **ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases**, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017* https://doi.org/10.1109/CVPR.2017.369

[3] Worawate Ausawalaithong, Sanparith Marukatat, Arjaree Thirach, Theerawit Wilaiprasitporn, **Automatic Lung Cancer Prediction from ChestX-ray ImagesUsingDeep Learning Approach**, *11th Biomedical EngineeringInternational Conference(BMEiCON)*, 2018 https://doi.org/10.1109/BMEiCON.2018.8609997

[4] Faiz Ahmad Khan, Tripti Pande, Belay Tessema, Rinn Song, Andrea Benedetti, Madhukar Pai1, Knut Lönnroth Claudia, M. Denkinger, **Computer-aided reading of tuberculosis chest radiography: moving the research agenda forward to inform policy**, *European Respiratory journal, 2017* https://doi.org/10.1183/13993003.00953-2017

[5] Kaiming He Xiangyu Zhang Shaoqing Ren Jian Sun**, Deep Residual Learning for Image Recognition**,*IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016*

[6] Justin Johnson Andrej Karpathy Li Fei-Fei, **DenseCap:Fully Convolutional Localization Networks for Dense Captioning**, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),2016* https://doi.org/10.1109/CVPR.2016.494

[7] Stefan Jaeger,Sema Candemir, Sameer Antani, Yì-Xiáng J. Wáng, Pu-Xuan Lu, and George Thoma, **Two public chest X-ray datasets for computer-aided screening of pulmonary diseases**, *Quantitative imaging in medicine and surgery, 2014*

[8] Guest Editorial Deep Learning in Medical Imaging**: Overview and Future Promise of an Exciting New Technique**, *IEEE Transactions On Medical Imaging, 2016*

[9] K. Simonyan and A. Zisserman**, Very deep convolutional networks For large-scale image recognition,** vol. abs/1409.1556, 2014.[Online].Available: http://arxiv.org/abs/1409.1556

[10] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, **Learning deep features for discriminative localization**, *IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), June 2016. https://doi.org/10.1109/CVPR.2016.319

.