2019 5th International Conference for Convergence in Technology (I2CT)
Pune, India. Mar 29-31, 2019

Early Prediction of Lung Diseases

Anuradha D. Gunasinghe
Informatics Institute of Technology,
57, Ramakrishna Rd, Colombo 06,
Sri Lanka.
Anuradhadenuwan@gmail.com

Achala C. Aponso
Informatics Institute of Technology,
57, Ramakrishna Rd, Colombo 06,
Sri Lanka.
Achala.a@iit.ac.lk

Harsha Thirimanna
Wso2
20,Palm Grove,
Colombo 3.
harsha.thirimanna@gmail.com

Abstract- **Machine learning is a branch of artificial intelligence that employs a variety of statistical, probabilistic and optimization techniques that allows computers to "learn" from past examples and to detect hard-to-discern patterns from large, noisy or complex data sets. Machine learning offers a principle approach for developing sophisticated, automatic, and objective algorithms for analysis of high-dimensional and multimodal biomedical data. Machine Learning plays an important role in medical systems. Earlier identification of diseases, we can be helped to detect earlier and more accurately, which can save many people as well as reduce the pressure on the system. Lung diseases are the one of the leading cause of death. The early identification and prediction of a lung diseases have become a necessity in the research, as it can facilitate the subsequent clinical management of patients. Machine Learning based decision support system provide the contribution to the doctors in their diagnosis decisions. Project considered about the breathing problems of patients as well as Asthma, Chronic Obstructive Pulmonary Disease (COPD), Tuberculosis, Pneumothorax and Lung cancer. Machine Learning and Deep Learning used to process data as well as create models for diagnosing patients. Combining the processing of patient information with data from chest X-rays, using CNN with the well-known pre-trained model, Caps Net network for data this form are the methods used for this project to identify the lung diseases. Initially studied and analyzed the data set, then apply Machine Learning and Deep Learning to predict that the patient has a lung disease or not. Project is a binary classification with input is patient's data (age, gender, chest X-ray images & view position) and output is found what the diseases is or not. The aim of the paper is to detect and diagnose the lung diseases as early as possible which will help the doctor to save the patient's life. This paper describes how lung diseases was predicted and controlled, using Machine Learning.**

Key words: Deep Learning, Lung Diseases, Machine Learning

I. INTRODUCTION

The pressure on health is on the increase, and the environment, climate, man's lifestyle, the public's increasing risk, the world is changing every day. In 2015, 56 million people are dead, of which 68 percent have slowly developed. On the top 10 list there are two diseases related to the lungs. Lung diseases leading cause of death there for we will focus on this article in lung-disease.

"When you breathe, in your lungs fill oxygen with air and distribute it to blood. Your body cells work and require oxygen to grow [2]. During a normal day, human breathe nearly 23,000 times [18]. People with lung disorders have had difficulty breathing. Millions of people in the United States have lung disease. All kinds of lung diseases together are the same number 3 killer in the United States. Lung diseases are many of the disorders of the lungs, such as asthma, COPD, influenza, pneumonia and tuberculosis, lung cancer and other respiratory problems. Some lungs can cause respiratory distress [1].

A. LUNG DISEASES

Tuberculosis is an infectious disease, caused in most cases by microorganisms called Mycobacterium tuberculosis. The microorganisms usually enter the body by inhalation through the lungs [4]. One of the most serious disease in the world is lung cancer [8].Moreover it can be totally cure using early detection. Er et al. [4]stated that Pneumonia is an inflammation or infection of the lungs most commonly caused by a bacteria or virus. Moreover Pneumonia can also be caused by inhaling vomit or other foreign substances. According to Er etal. [4], Asthma is a chronic disease characterized by recurrent attacks of breathlessness and wheezing. During an asthma attack, the lining of the bronchial tubes swell, causing the airways to narrow and reducing the flow of air into and out of the lungs. COPD is a preventable and treatable disease state characterised by airflow limitation that is not fully reversible [4]. Furthermore airflow limitation is usually progressive and is associated with an abnormal inflammatory response of the lungs to noxious particles or gases, primarily caused by cigarette smoking.

B. Structuring Technologies

This section states the popular data formatting technologies available for Lung Diseases, discussing their features.

(a) Machine Learning

Machine learning is a branch of artificial intelligence that employs a variety of statistical, probabilistic and optimization techniques that allows computers to "learn" from past examples and to detect hard-to-discern patterns from large, noisy or complex data sets [6]. Machine learning offers a principled approach for developing sophisticated, automatic, and objective algorithms for analysis of high-dimensional and multimodal biomedical data [19].

978-1-5386-8075-9/19/$31.00 ©2019 IEEE

(b) Deep Learning

Deep learning is making major advances in solving problems that have resisted the best attempts of the artificial intelligence community for many years. It has turned out to be very good at discovering intricate structures in high-dimensional data and is therefore applicable to many domains of science, business and government [12].

C. *Algorithmic Analysis (IT aspect)*

This is a problem with the new datasets that have never been fully modeled, therefore I do not want to get stuck and I come up with the following approach:

➢ Convolutional Neural Network

This is a powerful algorithm for processing image data like this. Given the huge data in the full dataset, this is indeed the appropriate method to apply, some parameters are considered and used as follows:

- Neural network architecture: Choose the appropriate architecture

A neural network architecture is introduced for incremental supervised learning of recognition categories and multidimensional maps in response to arbitrary sequences of analog or binary input vectors, which may represent fuzzy or crisp sets of features [4].

- Preprocessing parameters
- Fine tuning
- Spatial transformer

This is a differentiable module which applies a spatial transformation to a feature map during a single forward pass, where the transformation is conditioned on the particular input, producing a single output feature map [11].

- Training parameters

The training parameters are important because they emphasis the required performance and the accuracy required from the neural network [3].

- Add more data in network not only images

➢ **Capsules Network**

With the power to distinguish many objects from different perspectives, I find it useful because our image data has two types of View Position. Just like CNN I have some things to do as follows:

- Capsules Networks architect: Choose the appropriate architecture
- Preprocessing parameter
- Training parameters

II. METHODOLOGY

Recently a large dataset of X-ray lung data was public on Kaggle and UCI Machine Learning Repository followed by labeled lung disease data.

A literature survey was conducted to gain the background knowledge an learn all the related techniques and technologies like Machine Learning, Deep Learning , Convolutional Neural Networks(CNN),etc. It was mainly focus on the lung diseases prediction system. Optimized CNN is the main method of this project. Here is the Architecture for this approach.

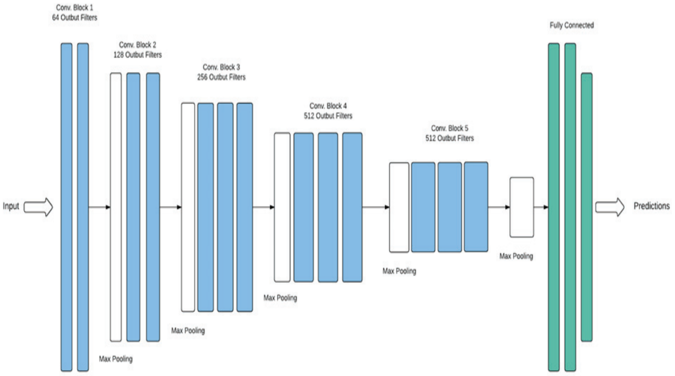

Fig. 1: VGG16 architect extract feature

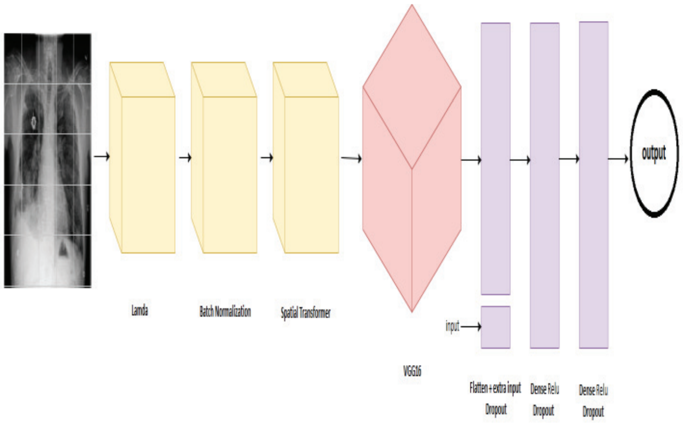

Fig. 2: Full architecture

The architecture consists of three main layers in the following order:

➢ Spatial transformer layers (The first three layers)

- The first is lamda to transfer the routing features [-0.5: 0.5], which means that the features of the image have an average value of 0.
- The second is Batch Normalization.
- The third layer is Spatial Transformer, which is used to extract the most valuable features for classification.

➢ Extract features layers (VGG16 pretrained model)

- A set of 13 layers as shown in the first image of the VGG is the extract features, there are many pretrained model but now I am trying before with VGG16 because this is a simple model for learning time and training faster.

➢ Classification layers (Last 3 layers)

- The first layer is the Flattened layer from the output of the layers VGG16 and 5 features plus 'Age', 'Gender Male', 'Gender Female', 'View position AP', 'View position PA'. These additional features will also affect the sorting, as we have seen above, so they are added to this layer. Following this layer is the dropout layer.
- The next two layers are Dense after each Dropout layer, with a gradual decrease in depth.

Capsule Network

With the Capsule Network I had a slight change from the original Hinton architecture so that it could work well with this data set. Here is the architecture taken from the article by Hinton, I will clarify the changes right below:

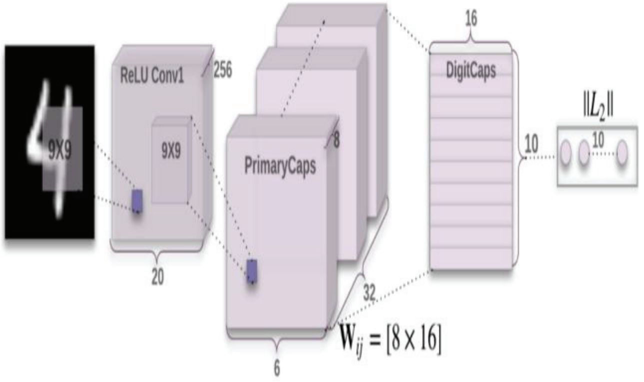

Fig. 3: Capsule Network

Optimized CNN

Changing and experimenting with a lot of image sizes, I found that the 64x64 image size was small enough and good enough for the model to capture the pattern of the image. Use the Spatial transformer with some layers supporting the front as lamda layer. The spatial transformer layer uses a fairly simple locnet (localization network) model to separate key features from the image. Non-complementary data has been tested in many places on the architecture, and the first layer of the classification is most appropriate. Tweaks the thresholds of precision, recall, and Fbeta score

Refine the index of the dropout layer in the classification Parameter of optimizer Gradient descent with momentum decay and learning rate.

III. PROBLEM IDENTIFICATION

People with lung disease have difficulty of breathing. Millions of people have lung disease in the U.S. If all types of lung disease are lumped together, it is the number three killer in the United States. The term lung disease refers to many disorders affecting the lungs. Now a days there is no computer system for the identification of the all lung diseases using chest X ray although there is a system for the identification of Pneumonia. This project is provided the diagnosis of lung diseases from the patient's chest X-ray data plus some additional information.

IV. OVERVIEW OF A POSSIBLE SOLUTION

This project proposed the diagnosis of lung diseases from the patient's chest X-ray data plus some additional information. The best solution is to have a complex CNN with the following data processed:

- Research for resolved issues, domain information, support data, methods, and solution data for similar projects. Some potential techniques are listed and investigated.

- Sample data is downloaded and analyzed, preprocessing, metric selection
- Testing multiple architectures, optimizing and testing on a sample dataset.
- Use good architects to test the full dataset, continue optimizing and statistics.

This project is based on a very new set of data and not many people find out, this is a very good problem and if done well it will make a big contribution to the community. This project has tested many new and interesting methods such as Spatial Transformer and has shown that they have recorded remarkable results.

This project is hardly new, and chest X-rays are difficult to see clearly, the data is not standardized, and NLP labeling can be used to obtain the disease. It is also difficult to apply a very new method of Convolutional Neural Network so there is not much documentation to optimize it. Big data on the full dataset is also a big challenge for me being limited to computer power.

The results of this project has achieved my initial expectation, but should be able to apply in hospitals, more improvements are needed to increase the precision of the model.

V. CONCLUSION AND FUTURE WORK

In this paper, the effect of the lungs of a modern patient on the various researchers and the damage to the lung is clearly explained by various researchers. Since these diseases (asthma, COPD, pneumonia, tuberculosis and lung cancer) has been cured the necessity of identifying this disease has become essential according to many researches as. One of the main concerns of this research is to identify and select a proper data sets and technique to analyze lung diseases. Chest x-ray was selected based on the comparisons and discussions that were stated in this paper.

Next a proper and suitable feature extraction algorithm was chosen since the chest x-ray signal may contain lots of unnecessary data. This selection was based on advantages and disadvantages of using many common algorithms, such as CNN, Capsule Network. Finally, a classification algorithm was also discussed based on their characteristic qualities. In short-term research, it was seen that CNN algorithm added additional benefits to predict the lung diseases in advance with better results. Ultimately, lung disease can be diagnosed.

In the future, we hope to conduct a training with more data sets and change some parameters to faster the model. Some metric parameters of the metrics will also be tested. We can experiment on pre-trained model to improve the accuracy.

REFERENCES

[1] Anon., 2018. Two lung diseases killed 3.6 million in 2015: study. [Online]
[Accessed 4 November 2018]

[2] Anon., 2007. Understanding how your lungs work and how COPD affects your body. s.l.:s.n.

[3] Barghash, M. A. & Santarisi, N. S., 2004. Pattern recognition of control charts using artificial neural networks - analyzing the effect of the training parameters. Journal of Intelligent Manufacturing, Volume 15, pp. 635-644.

[4] Carpenter, G. A. et al., 1992. Fuzzv ARTMAP: A Neural Network Architecture for incremental supervised learning of analog multidimensional maps. IEEE Transactions on Neural Network, 3(5).

[5] Celli, B. R. & MacNee, W., 2004. Standards for the diagnosis and treatment of patients with COPD: a summary of the ATS/ERS position paper. *European Respiratory Journal,* Volume 23, pp. 932-946.

[6] Cruz, J. A. & Wishart, D. S., 2006. Applications of machine learning in cancer prediction. Volume 2, pp. 59-78.

[7] Dept. of Health and Human Services Office on Women's Health :https://medlineplus.gov/lungdiseases.html

[8] Durga, S. & Kasturi, K., 2017. Lung disease prediction system using data mining techniques. Jour of Adv Research in Dynamical & Control Systems, 9(5).

[9] Er, O., Yumusak, N. & Temurtas, F., 2010. Chest diseases diagnosis using artificial neural networks. Expert Systems With Applications, Volume 37, pp. 7648-7655.

[10] https://www.womenshealth.gov/a-z-topics/lung-disease

[11] Jaderberg, M., Simonyan, K., Zisserman, A. & Kavukcuoglu, K., 2015. *Spatial Transformer Networks.* s.l., Neural Information Processing Systems Conference.

[12] LeCun, Y., Bengio, . Y. & Hinton, G., 2015. Deep learning. *International journal of science,* pp. 436-444.

[13] *Matrix capsules with EM routing.* https://openreview.net/forum?id=HJWLfGWRb&noteId=HJWLfGWRb

[14] Max Jaderberg, Karen Simonyan, Andrew Zisserman, Koray Kavukcuoglu. *Spatial Transformer Networks*. https://arxiv.org/abs/1506.02025

[15] Misra, A., Rudrapatna, M. & Sowmya, A., 2004. Automatic Lung Segmentation: A Comparison of Anatomical and Machine Learning Approaches.

[16] *NIH full Chest X-rays dataset*, https://www.kaggle.com/nih-chest-xrays/data

[17] *NIH sample Chest X-rays dataset*, https://www.kaggle.com/nih-chest-xrays/sample

[18] Niwa, H. (2007) '[No Title]', *Development*, 134(4), pp. 635–646.

[19] Sajda, P., 2006. Machine Learning for Detection, New York: s.n

[20] Sara Sabour, Nicholas Frosst, Geoffrey E Hinton. *Dynamic Routing Between Capsules.* https://arxiv.org/abs/1710.09829

[21] Van, D. D., 2018. Diseases detection from Chest X-ray data, s.l.: s.n.

[22] Vinitha, S., Sweetlin, S., Vinusha, H. & Sajini, S., 2018. Disease prediction using machine learning over big data. *Computer Science & Engineering: An International Journal (CSEIJ),* 8(1).

[23] Wang, X. et al., n.d. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases.