JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Thoracic Disease Detection Using Supervised Learning

*1RICHA TIWARI, 1Prof. MONIKA VERMA, 2Prof. SUMIT KUMAR SAR
*1 PG Student, 1 Assistant Professor, 2 Associate Professor
*1Department of Computer Science Engineering,
1Bhilai institute of Technology, Durg (491001), India

Abstract: Innovation in the sector of Machine learning in the medical field can help in diagnosis and treatment of several diseases with absolute accuracy. Implementation of machine learning in healthcare has already proven to be a positive experience for patients as it also contributes towards intense care. In our paper, we have worked on diagnosis of different thoracic disease using the Machine Learning model and Chest X-ray images as input. The main focus of our work was on the Supervised Machine Learning Model i.e, Logistic Regression, which was implemented with Principal Component Analysis, and we obtained the highest accuracy of 0.78.

IndexTerms - **Machine Learning, Logistic Regression, Principal Component Analysis, Chest X-ray, Thoracic Disease.**

I.INTRODUCTION

Machine learning is being increasingly commonly employed in healthcare, and it is assisting patients and physicians in a variety of ways. Automatic diagnostic medical image processing is one important activity that can help strengthen the health industry and save the lives of millions of people across the world. As a result, significant investment is being made in models that can increase the performance and precision of doctors as well as the treatment of patients. Detecting the so-called "edge-cases" is often necessary for identifying unusual or difficult-to-diagnose disorders. Because this type of machine learning system is based on enormous datasets comprising raw images (and numerous transformations) of these diseases, it is typically more reliable than people at detecting them. More importantly, researchers are employing machine learning (ML) to create lots of new intelligent solutions that will aid in the diagnosis and treatment of illnesses. Patients are the one who will gain the biggest benefit from the technology since it can enhance the chance of positive outcomes by analysing the best therapy options for them. Machine Learning can detect abnormalities with accuracy at its early stage saving several bills, risk and life. The technology today has been developed so much and continues to develop to the extent where it can identify and treat even the extremely complicated conditions of the human body.

In our model, we have worked with only six types of thoracic diseases and those diseases were randomly classified into three sets. Infiltration and Pneumothorax were in Set A, Pleural Thickening and Emphysema were in Set B and Fibrosis and Pneumonia were in Set C. Although there are many different forms of thoracic disorders that affect the lungs, chest and nearby organs causing difficulty in breathing, we focused on Infiltration, Pleural Thickening Fibrosis, Pneumothorax, Pneumonia, and Emphysema in our project.

We have used Chest X-ray images as input that goes through various training and testing processes to help in identification of abnormalities. Computer Aided Design (CAD) helps in running the smooth process since, CAD technique helps in extracting only important attributes that contribute towards accuracy in result, efficiently. We have Logistic Regression (supervised algorithm) with PCA implemented in our model that helps in maintaining the maximum attributes of the input while reducing the dimension, and all for better results.

II. EXISTING STUDIES

Several studies have been conducted in the medical field employing AI techniques to aid in the early diagnosis of diseases, which has benefited patients. Various disorders affecting the lungs, chest, heart, and respiratory system, among others, are dangerous in the human body and can be detected early using the CAD system. Here is some research which has been done in the past for diagnosis of thoracic disease.

Lung cancer is the leading reason behind cancer-related deaths all round the world. Subapriya V. et al. [1] within the year 2020 work on CAD for lung cancer prediction using Convolutional Neural Network (CNN) and ML approach. In this research, they remove some audio in the CT image dataset along with the background images and more. The information set is then formed from 60 anterior-poster (PA) chest x-rays images, collected from normal cases at the University Hospital of Santiago Radio Diagnostic Team. Later, feature extraction was performed using CNN and everything ended up predicting lung cancer, using computer aided diagnosis. Anuradha D. Gunasinghe et al. [2] have also worked on early detection of lung diseases. They focused on the breathing problems of patients and many other disorders, such as Asthma, Chronic Obstructive Pulmonary Disease (COPD), Tuberculosis, Pneumothorax and Lung Cancer. Using CNN with the pre-trained model, Caps Net Network, for this data type, was the method used in this project to determine the lung anomalies. The aim of the paper was to identify and diagnose lung diseases at the earliest so that it can help the doctors to save patient's lives. This paper describes how, using Machine learning, how lung diseases were predicted and controlled.

Since some lung diseases even cause respiratory failure which is why, Murat Aykanat et al. [3] compared different Machine learning models such as, support vector machine (SVM), k-nearest neighbor (k-NN) and gaussian bayes (GB) in the classification of respiratory diseases with text and audio data. An electronic stethoscope and its software were used to record patient information (17,930 lung sounds from 1630 individuals). In the classification of lung diseases, SVM with text data was the most accurate while in the classification via audio data, K-nn turned out as the foremost accurate. Using both audio and text data, SVM was absolutely accurate. When they classified healthy versus sick via text, audio and combined data, Gaussian Bayes was precise and efficient among all, closely followed by K-nn. Thus the researchers concluded that there are a large number of features but a limited number of samples. Apart from that they found, SVM and K-nn were best in the classification of the dataset. However, Gaussian Bayes was perfect among all when it involved classification into two classes. Pragya Chaturvedi et al. [4] focused on list, compare, discuss, and analyse several methods in feature extraction, image segmentation, and various techniques to categorise and identify lung cancer in their early stages. She aimed to list out all the main research that was done over the years in the past that can be improved to reach perfection in outcome.

Data mining techniques (association rule mining, classification and clustering) have been implemented to analyse the several diseases. S. Durga et al. [5] proposed the model to predict the disease at earliest based on the symptoms, and data processing techniques like classification and clustering turned out to be helpful for the same cause. In the paper, the disease was predicted by using the info-mining hybrid approach. In that kind of approach, the user must enter the symptoms associated with the disease they are affected by. The result will be generated as per the abnormalities and its level of infection after the process of mapping the symptoms of the user in the database was done. Benjamin Antin et al. [6] analysed the research paper that was based on the detection of pneumonia using the chest x-ray. The datasets were frontal chest x-rays and were labeled with 14 thoracic abnormalities such as pneumonia, fibrosis, etc., and no findings (in case the patient was healthy). The images that they used in their model were of 1024x1024 pixels which were resized using an anti-aliasing filter to the dimensions as per the need because the logistic regression baseline, which they have implemented, worked with 32x32 resolution. The AUC score of 0.60 was observed, which was lower than the expected. The reason discovered for low accuracy was the inability of logistic regression to capture the complexity of the massive database. Later, they implemented the Convolutional Neural Network model to overcome the drawbacks of previously used methods so that they can achieve better results. However, they achieved an AUC Score of 0.609 only. Wang, et al. [7] released an outsized dataset consisting of 108,948 frontal-view x-ray images from 32,717 different patients and implemented deep convolutional neural networks (DCNN). Shubhangi Khobragade et al. [8] proposed lung segmentation, lung feature extraction and its classification using artificial neural network (ANN) in feed forward direction for the detection of lung diseases like lung cancer, pneumonia and TB.

III. Model Implementation

Logistic Regression is the popular machine learning model and interestingly supervised algorithm in the field. It is mostly used as a supervised classification algorithm, and basically it's a binary classifier. In our project, we have worked on Principal Component Analysis (PCA) with Logistic Regression. PCA is a process where the dimension and variables of datasets are reduced while preserving maximum possible information. This process was introduced in our model as smaller datasets with maximum number of retained information is easier to explore and faster to work with.

Here in our model, we are extracting the images we are working with by resizing it from 1024x1024 pixel to 32x32 pixel along with Image flattening. Image flattening eases the exporting and printing process by significantly reducing the file size. After all these steps, PCA is applied on the resized outcome and then the images are transferred through the Logistic Regression model and finally through the Grid search that produces the best set of hyperparameters, which gives the best possible results. Grid search functions by passing all the combinations of hyperparameters in the model and checking the results simultaneously. Stepwise working of our model are as below:

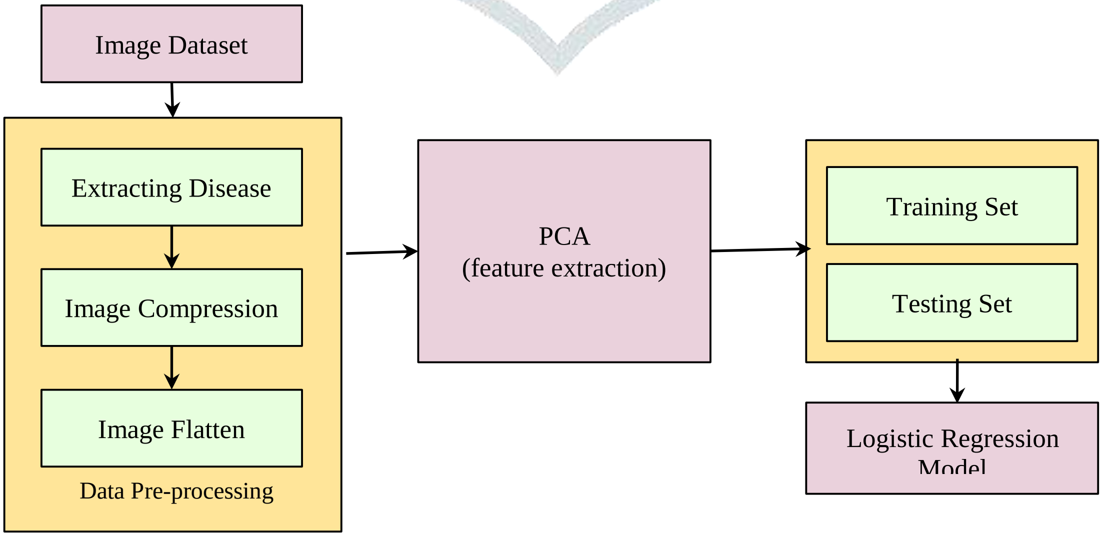

Figure-1: Model workflow

The first things we did was to connect our Google drive to 'colab' to save the preprocessed data and variables. Later 'colab' was connected to kaggle from where we downloaded the dataset for our model. After downloading the datasets, it was imported to the library and extracted wisely. Later, the dataset csv file that was extracted earlier, was processed under some minor preprocessing techniques. The dummy dataset was created after the preprocessing step that helped in separating diseases of each image. Now the image was displayed in its original and later form when the PCA was applied on it. Up to this point, most of the important features have been extracted from the image. Now, the number of total chest x-ray images were counted and the number of images with 'no findings' were determined. The image path was then added to the dataset and it was displayed while storing only the important columns, along with that the csv file was saved in the google drive to be used at some point after.

The further steps in our model included reading the csv file from the google drive to count the values of the dataset. The dataset was then separated into three groups where each group contained only two diseases at random. This step was crucial for efficient and accurate working of the model. After random classification of diseases in each group, both the diseases and groups were counted. By this time, we are working on a particular group that has two disease data and the same process will be applied on the other two remaining groups.

The first step will be resizing of the data that will convert pixels from 1024x1024 to 32x32, followed by the flattening process. Only the first 1024 columns will be stored and PCA will then be applied on the groups. The target and paths are then added to the dataset with label encoding being performed on the target column. After all these steps, Confusion Matrix code will be applied, which will be followed by the training and testing of the model with the first dataset and then by the other two dataset. The few final steps in our model include applying a confusion matrix on the first dataset followed by a scatter plot with PCA and marker label on the same dataset. To determine the final outcome from the entire model, the same working process will be applied on each and every dataset. Thus, we will be able to achieve the accuracy efficiently in the diagnosis of thoracic disease from the Chest X-ray images.

IV. Result

We have conducted several experiments in order to create a flawless model for diagnosis of Thoracic Disease whose description will be explained in this section. There are several kinds of thoracic disorders and we have worked upon six among all. We opted for Fibrosis, Pneumonia, Infiltration, Pneumothorax, Pleural Thickening and Emphysema containing chest x-ray images from the dataset. Our dataset containing six diseases were randomly divided into 3 sets namely A, B and C. Set A comprises Infiltration & Pneumothorax; Set B comprises Pleural Thickening & Emphysema; Set C comprises Fibrosis & Pneumonia.

At the beginning, we created a different combination of dataset containing all six diseases and a single model to work with in order to identify the disease, and it resulted in lower accuracy. Afterward, to achieve better results than the first attempt, we created two groups containing three diseases from our dataset and the results were almost the same as the initial one. Subsequently, we randomly divided our dataset into three groups (A, B, and C) where each group had two diseases and this attempt resulted with the higher accuracy.

The original images were in 1024x1024 pixels. In our working model, due to resource constraint, we had downscale the images into 32x32 pixels. Before reducing the image size into the lowest possible pixels, we tried reducing it into 128x128 pixels and 64x64 pixels but that caused system failure. Which is why, after the downsampling of images, we used a machine learning model called Logistic Regression with Principal Component Analysis (PCA) for extracting the maximum possible feature from the input images and excellent performance of the model. After that, we used the Hyperparameter tuning approach, also known as Grid Search, which assists in generating a model for each possible combination of all of the hyperparameter values provided, followed by the evaluation of the model, and determination of the model's design that can generate the better results.

Following the application of Logistic Regression with PCA on our entire dataset, we obtained the following 'Train and Test' accuracy:

- On the training and testing sets, Set A (Infiltration & Pneumothorax) computed accuracy of **0.748** and **0.726**, respectively.
- On the training and testing sets, Set B (Pleural Thickening & Emphysema) had an accuracy of **0.69** and **0.61**, respectively, and
- On the training and testing sets, Set C (Fibrosis & Pneumonia) had an accuracy of **0.78** and **0.76**, respectively.

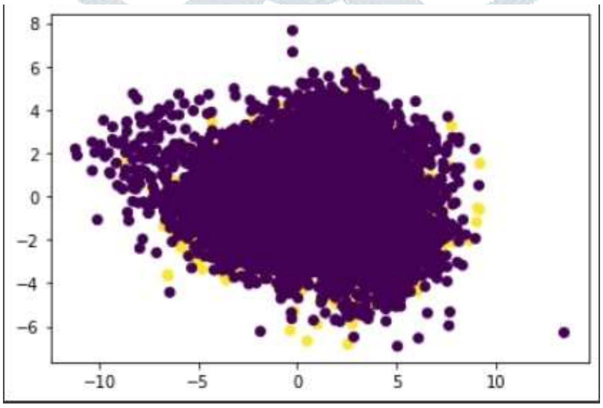

Figure-2: Scatter Plot for Database A

For Dataset A, Figure(2) displays scatterplot. Scatter plots are used to observe and display relationships between two numeric variables. In a scatter plot, the dots show the values of individual data points along with the patterns when the data is viewed as a whole.

The next example is a marker with both true and guess values.

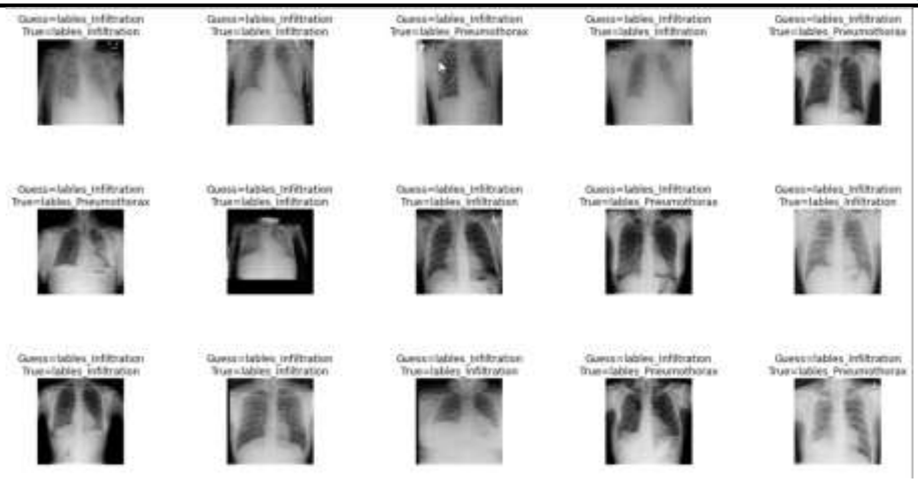
Guess=lables_Infiltration
True=lables_Infiltration
Guess=lables_Infiltration
True=lables_Infiltration
Guess=lables_Infiltration
True=lables_Pneumothorax
Guess=lables_Infiltration
True=lables_Infiltration
Guess=lables_Infiltration
True=lables_Pneumothorax
Guess=lables_Infiltration
True=lables_Pneumothorax
Guess=lables_Infiltration
True=lables_Infiltration
Guess=lables_Infiltration
True=lables_Infiltration
Guess=lables_Infiltration
True=lables_Pneumothorax
Guess=lables_Infiltration
True=lables_Infiltration
Guess=lables_Infiltration
True=lables_Infiltration
Guess=lables_Infiltration
True=lables_Infiltration
Guess=lables_Infiltration
True=lables_Infiltration
Guess=lables_Infiltration
True=lables_Pneumothorax
Guess=lables_Infiltration
True=lables_Pneumothorax

Figure-3: Prediction of actual value for input image for Dataset A.

We used performance indicators like F1 Score and ROC Accuracy to see how well the model performed for dataset A. The F1 score is 0.50, and the ROC accuracy is 0.52.

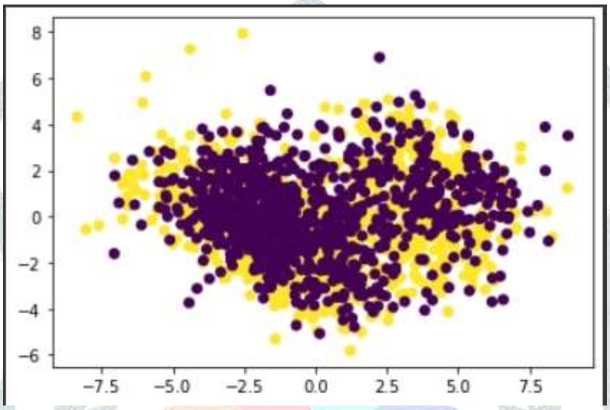

Figure-4: Scatter Plot for Database B.

For Dataset B, Figure(4) displays scatterplot. Scatter plots are used to observe and display relationships between two numeric variables. In a scatter plot, the dots show the values of individual data points along with the patterns when the data is viewed as a whole.

The next example is a marker with both true and guess values.

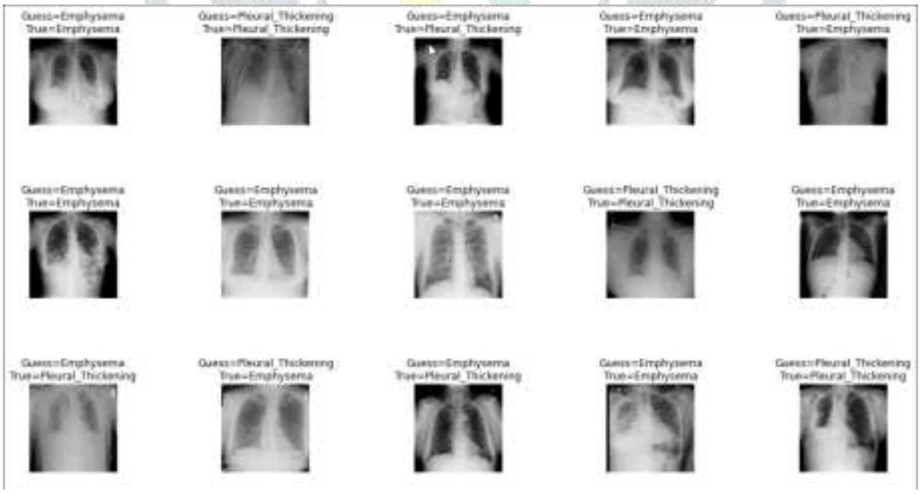

Figure-5: Prediction of actual value for input image for Dataset B.

We used performance indicators like F1 Score and ROC Accuracy to see how well the model performed for dataset A. The F1 score is 0.59, and the ROC accuracy is 0.60.

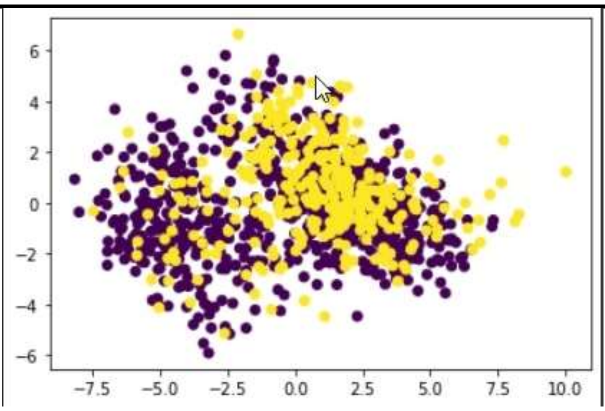

Figure-6: Scatter Plot for Database C.

For Dataset C, Figure(6) displays scatterplot. Scatter plots are used to observe and display relationships between two numeric variables. In a scatter plot, the dots show the values of individual data points along with the patterns when the data is viewed as a whole.

The next example is a marker with both true and guess values.

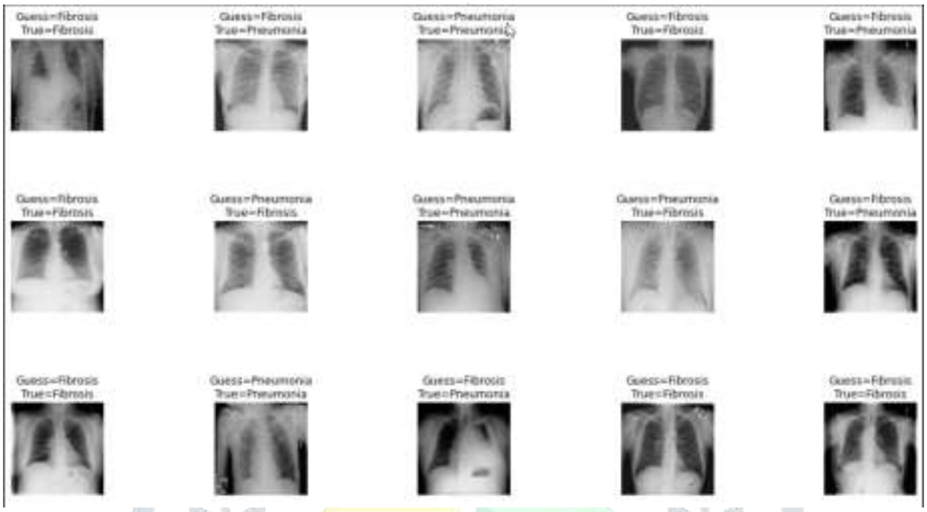

Figure-7: Prediction of actual value for input image for Dataset C.

We used performance indicators like F1 Score and ROC Accuracy to see how well the model performed for dataset A. The F1 score is 0.64, and the ROC accuracy is 0.63.

Table-1: Comparison table based on performance measure used (Logistic Regression with PCA).

Dataset	F1 Score	ROC accuracy
Dataset A (Infiltration & Pneumonia)	0.50	0.52
Dataset B (Pleural Thickening & Emphysema)	0.59	0.60
Dataset C (Fibrosis & Pneumonia)	0.64	0.63

V. Conclusion & Future Scope

In this paper, the types and effects of thoracic disease are clearly explained by various researchers. Our goal was to build a CAD system for efficient diagnosis of thoracic abnormalities. We have used the NIH Chest X-ray dataset containing 14 different kinds of thoracic diseases and have worked on 6 of them such as Fibrosis, Pneumonia, Infiltration, Pneumothorax, Pleural Thickening and Emphysema. All the selected dataset were downsampled from 1024x1024 pixels to 32x32 pixels. We divided the selected dataset into 3 groups containing 2 diseases in each group. Furthermore, a supervised learning technique, Logistic Regression was applied with PCA upon the datasets which were divided into the sets. Apart from this, various performance measures were used to see the performance of the model. However, due the resource constraint the overall result was not satisfactory or as expected.

Deep learning, particularly Convolutional Neural Networks (CNNs), will be used in the future to take an input image, assign importance (learnable weights and biases) to distinct aspects/objects in the image, and distinguish one from the other.

REFERENCES

[1] Subapriya V, Jaichandran R, Shunmuganathan K.L, Abhiram Rajan, Akshay T, Shibil Rahman, (2020) “ A Computer Aided Diagnosis of Lung Disease using Machine Learning Approach”, European Journal of Molecular & Clinical Medicine, Volume 7, Issue 4, pp 2662-2667.

[2] A. D. Gunasinghe, A. C. Aponso and H. Thirimanna, "Early Prediction of Lung Diseases," 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), 2019, pp. 1-4.

[3] AYKANAT, Murat & Kilic, Ozkan & KURT, Bahar & SARYAL, Sevgi. (2020). Lung disease classification using machine learning algorithms. International Journal of Applied Mathematics Electronics and Computers. 8. 125-132.

[4] Pragya Chaturvedi, Anuj Jhamb, Meet Vanani, Varsha Nemade, (2021), “ Prediction and Classification of Lung Cancer using Machine Learning Techniques”, IOP Conf. Series: Materials Science and Engineering.

[5] Durga, S. & Karuppiah, Kasturi. (2017). Lung disease prediction system using data mining techniques. Journal of Advanced Research in Dynamical and Control Systems. 9. 62-66.

[6] Benjamin Antin, Joshua Kravitz, and Emil Martayan. 2017. Detecting Pneumonia in Chest X-Rays with Supervised Learning. (2017).

[7] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. IEEE, 34623471

[8] Khobragade, Shubhangi & Tiwari, Aditya & Patil, C.Y. & Narke, Vikram. (2016). Automatic detection of major lung diseases using Chest Radiographs and classification by feed-forward artificial neural network. 1-5.

JETIR