Prediction Algorithm from Chest X-Rays of Thorax Disorders using Machine Learning

Sakshi[1], Srinivas Achar G[2]
[1, 2] Dept of Computer Science & Engineering
[2]Assistant Professor, Dept of Computer Science & Engineering
[1, 2] Atria Institute of Technology, Bangalore¬

Abstract- *The chest X-ray is among the top most commonly accessible medical imaging examinations used for affordable screening and diagnosis of numerous lung ailments including Pneumonia, Fibrosis, Hernia, Edema, Emphysema, Cardiomegaly and Pneumothorax. Owing to huge numbers of patients and increasing burden of lung ailments, the workload of radiologists has significantly multiplied. Hence, with an intention to accelerate/support the predictions of radiologists, machine learning methods can be implemented.A fundamental task in understanding CXR is to recognize the lung fields and the heart regions. The techniques such as Computerized image segmentation and feature analysis are useful for doctors in treating and diagnosing the disorder effectively.*

Keywords- Independent binary classifier, Logistic Regression, SIFT(Scale invariant feature transform), SVM(Support Vector Machines), Thoracic diseases, Visual bag of words.

I. INTRODUCTION

A substantial amount of time is spent by radiologists to find and detect the lung ailments by examining chest X-Ray. Chest X-Rays are the most common type of medical imaging, often 2x-10x more than other advanced imaging methods such as MRI, CT scans, PET scans. Inspection of chest X-ray is one of the most frequently used and cost-effective image examination process. The examination of X-Ray requires cautious observation and skills in the field of anatomy, physiology and pathological principles. The average time it takes a well-trained radiologist to read a CXR is about 1–2 minutes. It is hard to speed that up because CXR reading is a very systematic process.Incertain cases, doctors overlooked the ailments in the first examination on CXR, and when the images were re-examined, the disorder signs were discovered. An automated system can be beneficial to the patients who cannot afford the expertise.

The approach includes Machine Learning methods. The method is applied in building independent binary classifier for each of the disease. Grey scale images are taken into consideration. These images are pre-processed by resizing and cropping. Computer vision algorithm such as SIFT can be applied on pre-processed image which can detect feature descriptors in the images. The obtained feature descriptors can be used to construct visual bag of words. Machine learning techniques like logistic regression and SVM require a feature vector and computer visual ag of words can be used as feature vector.

II. LITERATURE SURVEY

Multiple research papers were observed. Few of them included image processing methods while others were about the artificial neural networks for prediction of disorder from chest X-Ray Images.

Emon Kumar Dey, Hossain Muhammad Muctadir [1] et al. presents a method for identification of abnormal mass tissue on digital x-ray. It involved the template matching technique for identifying mass tissue. This execution included DCT (Discrete Cosine Transform) based template matching that reduces the matching time.

Jie Chen [2] et al. adopted a new framework to enlarge the dataset gradually. Using the augmented dataset to train a CNN model for the thoracic disorder examination, they improved the performance of the model significantly. The further target is to integrate multiple images without labels gathered from different hospital to enhance the performance of the CNN models.

Shubhangi Khobragade [3] et al. introduced automated system to discover the lung ailments specifically for Tuberculosis, pneumonia and lung cancer with the help of chest radiographs. The outcome indicates that image pre-processing methods like histogram equalization, image segmentation generates effective output for the chest radiographs. Pattern recognition technique like feed forward artificial neural network is providing effective output.

Abhishek Sharma, Daniel Raju [4] et al. have detected the lung section by rib cage boundary identification. Otsu thresholding is used to separate the pneumonia cloud from the healthy lung in the lung region, still working on other techniques that can be used for thresholding the CXR images to generate improved outcome.

Zurina Muda, NoraidahSahari [5] et al. have shared an experience on segmenting the lung shape on CXR image. The segmentation method begins by identifying the lung edge using canny edge detection filters. To improve the edge detection, Euler number method is applied. Later, morphology technique is used to make the lung edge better so that the final output of lung region can be generated.

III. CASE STUDY

The mechanism focuses on feature extraction methods such as Scale Invariant Feature Transform, classification machine learning algorithms such as Logistic Regression, Support Vector Machines and Computer Vision algorithm like Visual bag of words to assist in the prediction of lung disorders from chest X-Ray Images.

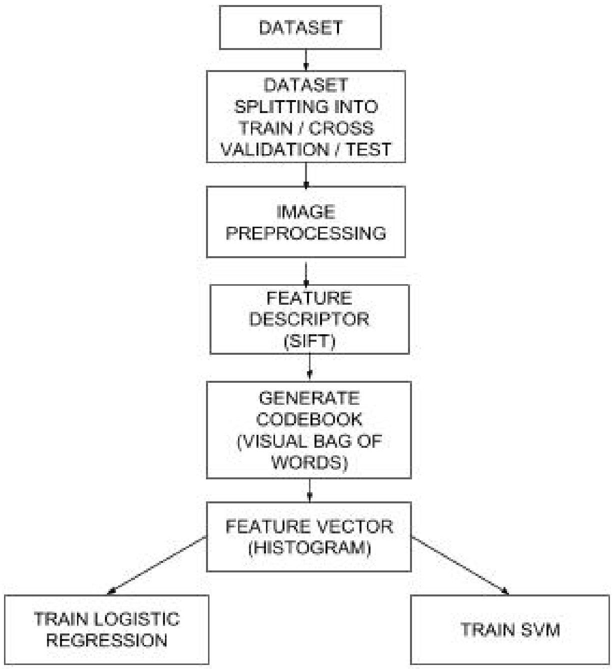
DATASET
DATASET SPLITTING INTO TRAIN / CROSS VALIDATION / TEST
IMAGE PREPROCESSING
FEATURE DESCRIPTOR (SIFT)
GENERATE CODEBOOK (VISUAL BAG OF WORDS)
FEATURE VECTOR (HISTOGRAM)
TRAIN LOGISTIC REGRESSION
TRAIN SVM

Fig. 1 Proposed system architecture

The dataset considered here is published by National Institutes of Health (NIH) Clinical Centre containing 100,000 plus frontal-view X-ray images of 30,805 unique patients composed of 14 lung disorders. Every image has multi-label images which are grey scale of size 1024 x 1024 in resolution. Data pipeline is used for splitting and pre-processing the data with the help of data split and pre-processing pipeline performance. The aim of pre-processing is an improvement of the image data that suppresses unwanted distortions or enhances some image features important for further processing. All the images will be resized from the resolution 1024 x 1024 to 224 x 224 to increase the computation speed. Then the images are cropped to make lungs in the image focal which results in the image with resolution of 180 x 200. Image contrast will be enhanced by applying the histogram equalizer. Histogram equalization is a method in image processing of contrast adjustment using the image's histogram. This method usually increases the global contrast of many images, especially when the usable data of the image is represented by close contrast values. The images will be split into training set, Cross-validation set and Test-set. Training dataset is the sample of data used to fit the model. Validation dataset is the sample of data used to provide an unbiased evaluation of a model fit on the training dataset while tuning model hyper parameters. The evaluation becomes more biased as skill on the validation dataset is incorporated into the model configuration. Test dataset is the sample of data used to provide an unbiased evaluation of a final model fit on the training dataset.

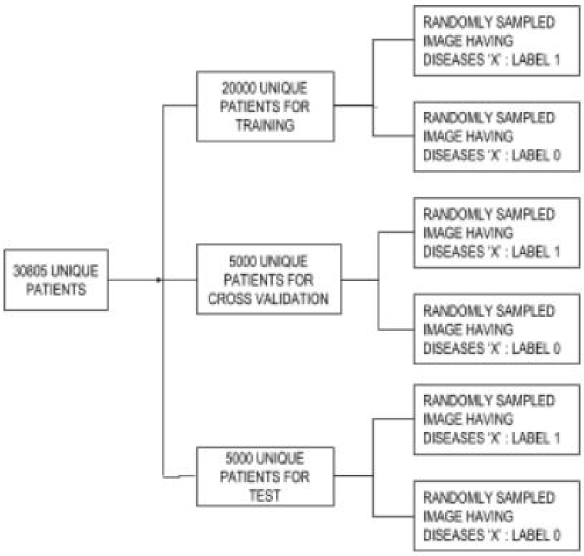
30805 UNIQUE PATIENTS
20000 UNIQUE PATIENTS FOR TRAINING
5000 UNIQUE PATIENTS FOR CROSS VALIDATION
5000 UNIQUE PATIENTS FOR TEST
RANDOMLY SAMPLED IMAGE HAVING DISEASES 'X' : LABEL 1
RANDOMLY SAMPLED IMAGE HAVING DISEASES 'X' : LABEL 0
RANDOMLY SAMPLED IMAGE HAVING DISEASES 'X' : LABEL 1
RANDOMLY SAMPLED IMAGE HAVING DISEASES 'X' : LABEL 0
RANDOMLY SAMPLED IMAGE HAVING DISEASES 'X' : LABEL 1
RANDOMLY SAMPLED IMAGE HAVING DISEASES 'X' : LABEL 0

Fig. 2 Data splitting for differentiating the lung disorders.

As every disorder will have independent binary classifier, discrete dataset will be generated for every disorder classifier. It is a type of supervised learning, a method of machine learning where the categories are predefined, and is used to categorize new probabilistic observations into said categories. Images will be sampled randomly for randomly sampled patients. Scale invariant feature transform is implemented to gain specific details in the image. SIFT finds the key points within an image and then calculates descriptor vector (refer figure no. 3) for each keypoint. Image is convolved with Gaussian filters at different scale, and then the difference of successive Gaussian blurred images is computed. Keypoints are the maxima or minima of the Difference of

Gaussian (DoG) that occurs at multiple scales. Orientation is computed for each keypoints (refer figure no. 4) based on local image gradient directions. Using orientation, descriptor vector is computed for each keypoint.

In computer vision, the bag-of-words model (BoW model) can be applied to image classification, by treating image features as words. A huge vocabulary of visual words is built using the bag of visual words which is also known as the codebook. Feature extraction is done with the help of SIFT, then will be generated followed by histogram. In order to obtain the codebook, the K-Means clustering is implemented for feature extraction for each image.Feature extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations. Feature extraction is related to dimensionality reduction. K-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. Every feature extraction is mapped to one of the nearest centroids. The count of features for each of the visual code words is given by the histogram of each image. For the training of the models, histogram is used as a feature vector.

For classifying, Logistic regression and SVM are the machine learning algorithms that will be implemented on visual bag of words feature vector to determine whether a chest X-ray is normal or infected with any disorder.

The simple linear regression model assumes that the linear relationship exists between the input and the output variables. The support vector machine searches the closest points and is known as “support vectors”. The name is as a result of the actual fact that points are like vectors which the simplest line “depends on” or is “supported by” the nearest points.

IV. EXPERIMENTATION

The execution of the process is carried out by pre-processing the images by cropping and resizing them so that the zone of concern can be rectified accurately. In order to discover the feature descriptors in the images, the SIFT computer vision algorithm will be implemented on the pre-processed images. These feature descriptors are used to build the visual bag of words. Computed visual bag of words will be used as a feature vector for Logistic regression and SVM. The output of each model will be a binary label for prediction of each disorder namely Pneumonia, Fibrosis, Hernia, Edema, Emphysema, Cardiomegaly, Pneumothorax.

The dataset of images consists of reasonable number of front-view X-rays of multiple unique patients, with every X-Ray labelled with up to 14 lung disorders. All the images in the dataset are grey scale images with the resolution of 1024 x 1024.

The metrics considered to evaluate the performance of models are accuracy, precision, recall and ROC (Receiver Operating Characteristic) curves. The number of cluster centroids for every classifier is deduced using accuracy and recall with more importance given to recall, because of the medical domain.

V. CONCLUSION

Logistic regression is implemented in the same manner as the SVM. By building the classifiers using conventional machine learning approach of drawing out features using computer vision technique, feasible performance is attained.

REFERNCES

[1] E. K. Dey and H. M. Muctadir, “Chest X-rayanalysis to detect mass tissue in lung,” *2014 International Conference on Informatics, Electronics & Vision (ICIEV)*, 2014.

[2] 2.J. Chen, X. Qi, O. Tervonen, O. Silven, G. Zhao, and M. Pietikainen, “Thorax disease diagnosis using deep convolutional neural network,” *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*,2016.

[3] S. Khobragade, A. Tiwari, C. Patil, and V. Narke, “Automatic detection of major lung diseases using Chest Radiographs and classification by feed-forward artificial neural network,” *2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES)*, 2016.

[4] A. Sharma, D. Raju, and S. Ranjan, “Detection of pneumonia clouds in chest X-ray using image processing approach,” *2017 Nirma University International Conference on Engineering (NUiCONE)*, 2017.

[5] M. N. Saad, Z. Muda, N. S. Ashaari, and H. A. Hamid, “Image segmentation for lung region in chest X-ray images using edge detection and morphology,” *2014 IEEE International Conference on Control System, Computing and Engineering (ICCSCE 2014)*, 2014.

[6] Rushikesh Chavan, Jidnasa Pillai*, Shravani Holkar, PrajyotSalgaonkar, Prakash Bhise , “Thoracic Diseases

Prediction Algorithm From Chest X-Ray Images Using Machine Learning Techniques”.

[7] https://www.wikipedia.org/