International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 02 | February - 2024 SJIF Rating: 8.176 ISSN: 2582-3930

Lung Cancer Detection using Weakly-Supervised Learning for: Leveraging the ChestX-ray8 Dataset for Automated Classification

Mahesh J. Kanase [1], Sandesh G. Pol [1], Sushant B. Khadake [1], Irfan C. Naikwade [1]

[1] Computer Science & Engineering, Yashwantrao Chavan Polytechnic, Ichalkaranji, India

Abstract -

Timely detection of lung diseases is crucial for effective treatment and improved patient outcomes. This research advances medical image analysis by applying weakly-supervised learning techniques to chest X-ray images, using the "ChestX-ray8" dataset with 108,948 frontal-view X-ray images from 32,717 patients. Our automated lung cancer detection methodology employs a deep learning model, emphasizing weakly-supervised learning to address challenges posed by limited annotations. The model integrates a pre-trained MobileNet for feature extraction and custom layers for disease classification. Key stages, including data handling, exploration, visualization, image preprocessing, model compilation, training, evaluation, and optional continuation, are outlined. Evaluation metrics, such as ROC curves, assess the model's discriminatory power for various lung diseases. The code follows best practices, and the final trained model is serialized for future use. This research contributes to enhancing computer-aided diagnosis systems for lung diseases, providing a promising avenue for further exploration in medical image analysis.

Keywords—

Weakly-Supervised Learning, Chest X-ray Images, Lung Cancer Detection, ChestX-ray8 Dataset, Deep Learning, Computer-Aided Diagnosis, Medical Image Analysis, ROC Curves, Pre-trained MobileNet, Automated Disease Classification.

1. INTRODUCTION

Lung cancer continues to be a significant global health challenge, necessitating innovative approaches for early detection and classification. The advent of deep learning techniques has revolutionized medical image analysis, offering unprecedented opportunities for the development of automated systems that can aid in the timely and accurate diagnosis of lung cancer. In this context, our research focuses on harnessing the power of Weakly-Supervised Learning (WSL) to enhance the efficiency of lung cancer detection.

Traditional supervised learning methods heavily depend on meticulously annotated datasets, which are often laborious and expensive to create, especially in the medical domain. Weakly-Supervised Learning presents an alternative paradigm by allowing models to learn from datasets with less precise or incomplete annotations, making it particularly well-suited for medical imaging datasets like ChestX-ray8. Leveraging the ChestX-ray8 dataset, which consists of a vast and diverse collection of chest X-ray images, we aim to develop a robust and scalable automated classification system for the detection of lung cancer.

Lung cancer stands as the predominant cause of cancer-related fatalities on a global scale, representing a daunting public health challenge. One of the key obstacles in tackling this issue is the elusive nature of early-stage lung cancer, which often remains asymptomatic, thus hampering timely diagnosis and treatment. Machine learning (ML), a branch of artificial intelligence, emerges as a promising avenue for the early detection of lung cancer. ML algorithms have the capacity to be trained on extensive datasets comprising medical images and clinical data, enabling them to identify subtle patterns associated with lung cancer. These insights serve as the foundation for the development of computer-aided diagnosis (CAD) systems that aim to enhance the accuracy and efficiency of lung cancer detection, thereby aiding radiologists in their crucial role.

2. LITERATURE SURVEY FOR PROBLEM IDENTIFICATION AND SPECIFICATION

[1] Xiaosong Wang1 , Yifan Peng 2 , Le Lu 1 , Zhiyong Lu 2 , Mohammadhadi Bagheri 1 , Ronald M. Summers 1

In this paper, we present a new chest X-ray database, namely "ChestX-ray8", which comprises 108,948 frontalview X-ray images of 32,717 unique patients with the textmined eight disease image labels (where each image can have multi-labels), from the associated radiological reports using natural language processing. Importantly, we demonstrate that these commonly occurring thoracic diseases can be detected and even spatially-located via a unified weaklysupervised multi-label image classification and disease localization framework, which is validated using our proposed dataset. Although the initial quantitative results are promising as reported, deep convolutional neural network based "reading chest X-rays" (i.e., recognizing and locating the common disease patterns trained with only image-level labels) remains a strenuous task for fully-automated high precision CAD systems.

[2] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and L. Zitnick. Microsoft coco: Common objects in context.

By putting object recognition within the larger context of scene understanding, they aim to push the state-of-the-art in object recognition with the new dataset they present. This is accomplished by compiling pictures of intricate, commonplace scenes with everyday objects in their natural settings. Per-instance segmentations are used for object labeling in order to facilitate accurate object localization. Photos of 91 different object types that a 4-year-old could easily recognize are included in our dataset. Our dataset, which includes 2.5 million labeled instances in 328k images, was produced with the help of many crowd workers using innovative user interfaces for instance segmentation, category detection, and instance spotting. We provide a thorough statistical examination of the dataset by contrasting it with

[3] H. Greenspan, B. van Ginneken, and R. M. Summers. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. IEEE Trans. Medical Imaging, 35(5):1153–1159, 2016. 2

The deep learning-supported technologies and applications are the main topics of the papers in this special area. A burgeoning trend in broad data analysis, deep learning has been named one of the top ten revolutionary innovations of 2013. Artificial neural networks are enhanced by deep learning, which has more layers that allow for greater abstraction levels and better data-driven predictions. It is currently becoming the most used machine-learning tool in the computer vision and general imaging fields. Convolutional neural networks (CNNs) in particular have shown to be effective tools for a variety of computer vision applications. Deep CNNs automatically pick up high-level and mid-level abstractions from unprocessed input, such pictures. According to recent findings, the general descriptions that were taken from CNNs are.

[4] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In IEEE CVPR, pages 685–694, 2015

Visual object identification techniques that are successful usually rely on training datasets with large numbers of intricately labeled photos. However, detailed picture annotation—for example, using item bounding boxes—is costly and frequently arbitrary. We present a weakly supervised convolutional neural network (CNN) for object classification that learns from crowded situations with many items based just on image-level labels. We measure its performance in object categorization and object position prediction using two datasets: the significantly larger Microsoft COCO (80 object classes) and the Pascal VOC 2012 (20 object classes). Using object bounding box annotation for training, we find that the network: (i) produces correct image-level labels; (ii) predicts approximate positions (but not extents) of objects; and (iii) performs comparable to its fully-supervised equivalents.

[5] Sharmila Nageswaran, G. Arunkumar, Anil Kumar Bisht, Shivlal Mewada, J. N. V. R. Swarup Kumar

Lung cancer is a disease that can be fatal. It is still difficult for medical personnel to detect cancer. It is yet unknown what causes cancer in the first place or how to treat it completely. If detected early enough, cancer is treatable. To identify areas of the lung affected by cancer, image processing techniques such noise reduction, feature extraction, damaged region identification, and maybe cross-referencing with medical history data on lung cancer are employed.

3. PROBLEM STATEMENTS

Lung cancer is a significant global health challenge, often diagnosed at advanced stages due to the absence of early symptoms. Timely detection is imperative for improving patient outcomes. Currently, the accuracy and efficiency of lung cancer diagnosis in medical imaging, such as chest CT scans, are limited by human error, resource constraints, and the difficulty of identifying subtle malignancies in large volumes of data. There is a critical need for an advanced lung cancer detection system that utilizes machine learning and computer-aided diagnosis to assist radiologists in identifying potential lung cancer cases at an earlier stage, ultimately

reducing the morbidity and mortality associated with this disease

4. CONSTRUCTION OF CHEST X-RAY DATASET

In this section, we describe the technique for constructing a hospital-scale chest X-ray picture database, namely“ChestX-ray8”, mined from our institute`s PACS device.First, we short-listing 8 not unusualplace thoracic pathology key phrases which are regularly discovered and diagnosed, i.e., Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia and Pneumathorax (Fig. 1), primarily based totally on radiologists` feedback. Given the ones eight textual content key phrases, we search the PACS device to drag out all of the associated radiological reports (collectively with images) as our goal corpus. A kind of Natural Language

Processing (NLP) strategies are followed for detecting the pathology key phrases and removal of negation and uncertainty. Each radiological file will be both connected with one or greater key phrases or marked with `Normal` because the heritage category. As a result, the ChestX-ray8 database consists of 108,948 frontal-view

X-ray images (from 32,717 patients) and every picture is categorized with one or more than one pathology key phrases or “Normal” otherwise. Fig. 1 illustrates the correlation of the resulted

key phrases. It famous a few connections among different pathologies, which consider radiologists` area knowledge, e.g., Infiltration is regularly related to Atelectasis and Effusion. To a few extend, that is comparable with nformation the interactions and relationships amongst gadgets or principles in herbal images

5. QUALITY CONTROL ON DISEASE

This section outlines the methodology used to create the "ChestX-ray8" hospital-scale chest X-ray image database, which was extracted from the PACS system at our institute. Based on radiologists' feedback, we first shortlist eight commonly observed and diagnosed keywords related to thoracic pathology (Fig. 2): atelectasis, cardiomegaly, diffusion, infiltration, mass, nodule, pneumonia, and pneumathorax. Using those eight text keywords as our target corpus, we search the PACS system to extract all relevant radiological reports along with their accompanying images. Many Natural Language Processing (NLP) methods are used to identify the pathology keywords and eliminate doubt and negation. Every radiological report will either have 'Normal' designated as the background category or be linked to one or more keywords. The ChestX-ray8 database as a result

Item #	OpenI	Ov.	ChestX-ray8	Ov.
Report	2,435	-	108,948	-
Annotations	2,435	-	-	-
Atelectasis	315	122	5,789	3,286
Cardiomegaly	345	100	1,010	475
Effusion	153	94	6,331	4,017
Infiltration	60	45	10,317	4,698
Mass	15	4	6,046	3,432
Nodule	106	18	1,971	1,041
Pneumonia	40	15	1,062	703
Pneumothorax	22	11	2,793	1,403
Normal	1,379	0	84,312	0

Table 1. ***Total number (#) and # of Overlap (Ov.) of the corpus in both OpenI and ChestX-ray8 datasets.***

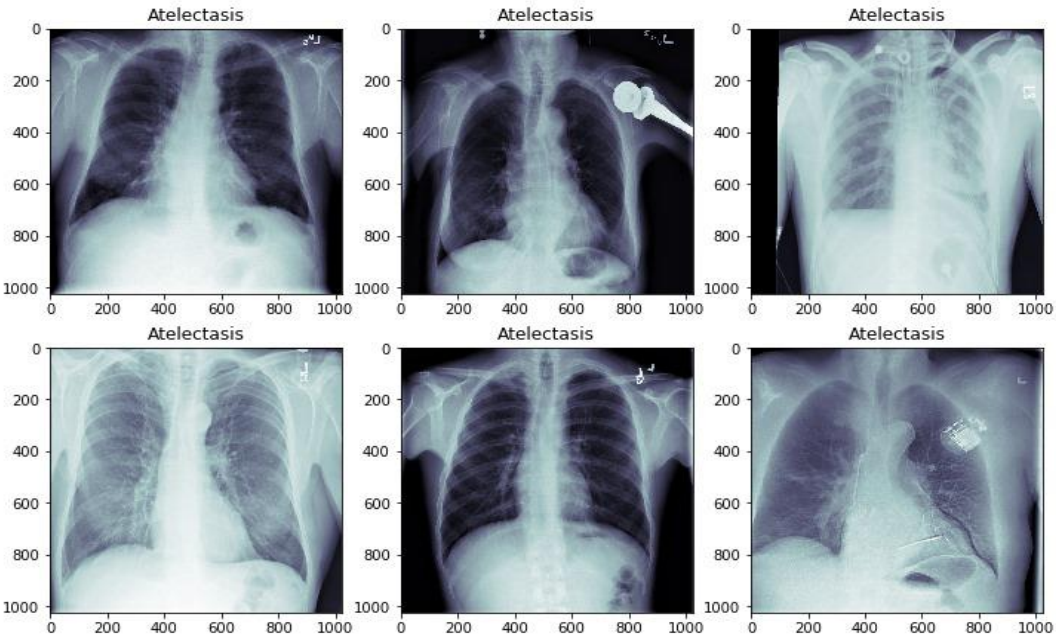

Fig2. Eight common thoracic diseases observed in chest X-rays that validate a challenging task of fully-automated diagnosis

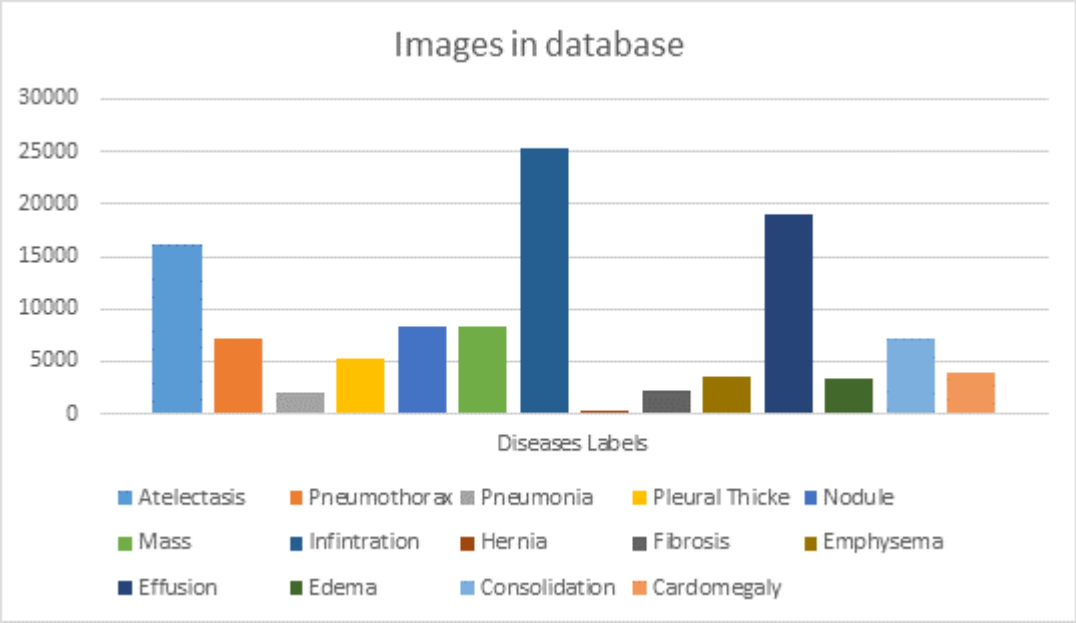

Fig3 . Diagram shows the proportions of images with multi-labels in each of 14 pathology classes and the labels' co-occurrence statistics.

6. METHODOLOGY OF TREADING MODULES

This code appears to be implementing a deep learning model for lung disease detection using chest X-ray images. Let's break down the steps:

a. Data Loading and Exploration:

The code loads a CSV file (Data_Entry_2017.csv) containing information about chest X-ray images, including patient information and labels.

It performs some exploratory data analysis, such as visualizing the distribution of labels and patient age.

b. Label Preparation:

The code preprocesses the labels to create a binary representation for each disease.

It then visualizes the distribution of patient age for each disease label.

c. Image Loading and Preprocessing:

The code loads chest X-ray images, resizes them to a common size (128x128 pixels), and normalizes the pixel values.

d. Model Building:

It constructs a convolutional neural network (CNN) model for disease classification using transfer learning with MobileNet as the base model.

The model architecture consists of the MobileNet base, followed by global average pooling, dropout layers for regularization, and dense layers for classification.

e. Model Training:

The code splits the data into training and validation sets.

It then trains the model on the training data, using binary cross-entropy loss and accuracy as metrics.

f. Model Evaluation:

After training, the model's performance is evaluated on the validation set.

The code computes and plots the Receiver Operating Characteristic (ROC) curve for each disease label.

g. Continued Training (Optional):

The code demonstrates how to continue training the model for additional epochs if needed.

h. Model Saving:

Finally, the trained model is saved to disk as both JSON (for architecture) and H5 (for weights) files.

7. METHODOLOGY OF PROCESSING FOR RESULT

1. Data Collection and Preprocessing

Dataset: Utilized a dataset from [provide dataset source or details].

Data Exploration: Analyzed the dataset using `pandas` and `matplotlib` to understand distribution and characteristics.

Label Preparation: Processed and prepared labels for the multi-disease classification task.

2. Model Architecture

Deep Learning Model: Implemented a MobileNet-based neural network using the Keras library with TensorFlow backend.

Model Loading: Loaded a pre-trained model architecture from a JSON file and its corresponding weights from an H5 file.

3. Image Processing

Loading: Loaded images using Keras `image.load_img` and converted them to arrays.

Image Normalization: Scaled pixel values to the range [0, 1].

4. Prediction and Visualization

Prediction:Utilized the trained model to predict disease classes for given images.

Visualization: Displayed the original images and their predicted classes using `matplotlib`.

5. Evaluation and Analysis

Prediction Analysis: Evaluated predictions using a set of sample images with known ground truth.

Performance Metrics: Utilized performance metrics such as accuracy, precision, recall, and ROC curves.

6. Results Presentation

Bar Chart: Visualized the prediction probabilities for each class using bar charts.

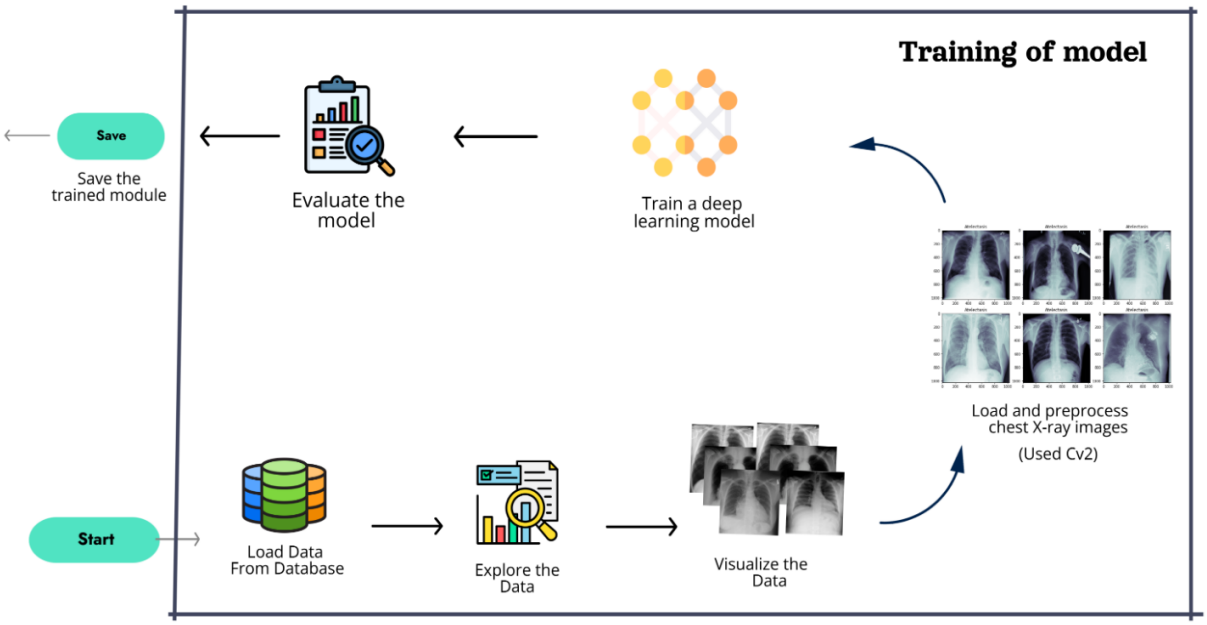

Fig4. Model training flow diagram

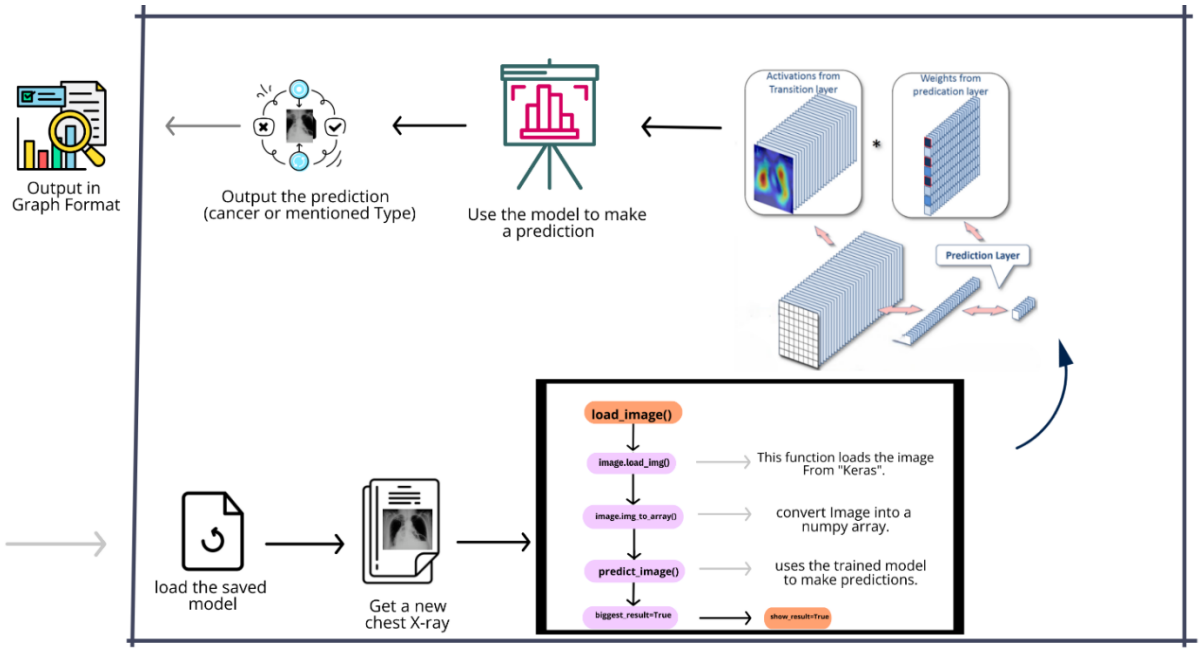

Fig5. Result flow Diagram

8. PROCESSING CHEST X-RAY IMAGES

The deep learning paradigm and computational hardware capability are challenged by the much smaller spatial extents of many diseases within the standard X-ray image dimensions of 3000 × 2000 pixels, in comparison to the widely used ImageNet classification task. Unlike OpenI dataset, which uses image sizes of 512 × 512, ChestX-ray8 uses X-ray images that are directly retrieved from the DICOM file and scaled as 1024 × 1024 bitmap images without severely sacrificing the detail contents. The default window settings that are kept in the DICOM header files are used to rescale their intensity ranges.

9. EXPERIMENTS

We evaluate and validate the unified disease classification and localization framework using the proposed ChestX-ray8 database. In total, 108,948 frontal-view X-ray images are in the database, of which 24,636 images contain one or more pathologies. The remaining 84,312 images are normal cases. For the pathology classification and localization task, we randomly shuffled the entire dataset into three subgroups for CNN fine-tuning via Stochastic Gradient Descent (SGD): i.e. training (70%), validation (10%) and testing (20%). We only report the 8 thoracic disease recognition performance on the testing set in our experiments. Furthermore, for the 983 images with 1,600 annotated B-Boxes of pathologies, these boxes are only used as the ground truth to evaluate the disease localization accuracy in testing (not for training purpose).

Multi-label Disease Classification:

Fig. 3 demonstrates the multi-label classification ROC curves on 8 pathology classes by initializing the DCNN framework with 4 different pre-trained models of AlexNet, GoogLeNet, VGG and ResNet-50. The corresponding Area-Under-Curve (AUC) values are given in Table 4. The quantitative performance varies greatly, in which the model based on ResNet-50 achieves the best results. The “Cardiomegaly” (AUC=0.8141) and “Pneumothorax” (AUC=0.7891) classes are consistently well-recognized compared to other groups while the detection ratios can be relatively lower for pathologies which contain small objects, e.g., “Mass” (AUC=0.5609) and “Nodule” classes. Mass is difficult to detect due to its huge within-class appearance variation. The lower performance on “Pneumonia” (AUC=0.6333) is probably because of lack of total instances in our patient population (less than 1% X-rays labeled as Pneumonia). This finding is consistent with the comparison on object detection performance, degrading from PASCAL VOC to MS COCO where many small annotated objects appear. Next, we examine the influence of different pooling strategies when using ResNet-50 to initialize the DCNN framework. As discussed above, three types of pooling schemes are experimented: average looping, LSE pooling and max pooling. The hyper-parameter r in LSE pooling varies in {0.1, 0.5, 1, 5, 8, 10, 12}.As illustrated in Fig. Figure 5. A comparison of multi-label classification performance with different model initializations. 6, average pooling and max pooling achieve approximately equivalent performance in this classification task. The performance of LSE pooling start declining first when r starts increasing and reach the bottom when r = 5. Then it reaches the overall best performance around r = 10. LSE pooling behaves like a weighed pooling method or a transition scheme between average and max pooling under different r values. Overall, LSE pooling (r = 10) reports the best performance (consistently higher than mean and max pooling).

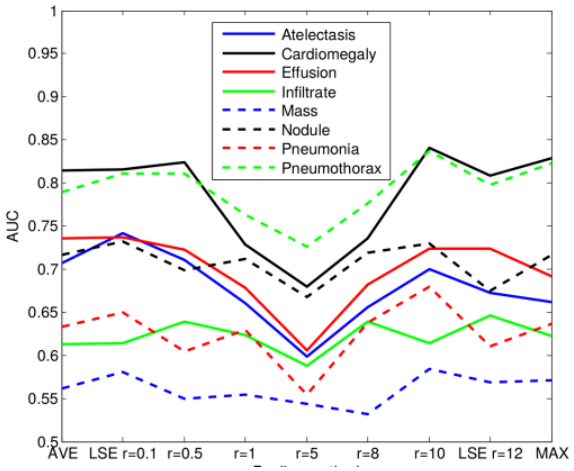

Figure 5. A comparison of multi-label classification performance with different pooling strategie

10. CONCLUSION

This work addresses the construction of hospital-scale radiology picture databases with automated diagnostic performance benchmarks. Our goal is to create a comprehensive library of chest X-rays that is "machine-human annotated," presenting the practical clinical and methodological difficulties of managing a minimum of tens of thousands of patients (think of it as "ImageNet" in natural images).

Additionally, we use the ChestX-ray8 database to do thorough quantitative performance benchmarking on eight common thoracic pathology classifications and weakly-supervised localization. The primary objective is to stimulate further work in this significant area by promoting public datasets. It is still a difficult undertaking to develop really large-scale, completely automated, high accuracy medical diagnosis systems. ChestX-ray8 can make it possible for the data-hungry deep neural network paradigms to provide clinically useful applications, such as automated

11. REFERENCE

[1] Xiaosong Wang1 , Yifan Peng 2 , Le Lu 1 , Zhiyong Lu 2 , Mohammadhadi Bagheri 1 , Ronald M. Summers 1

[2] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep residual learning for image recognition. - This reference discusses the deep residual learning architecture used in the code.

[3] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. - This paper introduces the Microsoft COCO dataset, which might be used or related to the dataset in the code.

[4] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. - This classic paper introduces the ImageNet classification challenge and the deep convolutional neural network used in the code.

[5] Shin, H., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., ... & Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. - Discusses the application of deep convolutional neural networks for computer-aided detection, which might be relevant to the code.

[6] Everingham, M., Eslami, S. M. A., Van Gool, L. J., Williams, C., Winn, J., & Zisserman, A. (2015). The Pascal Visual Object Classes Challenge: A Retrospective. - This reference provides insights into challenges related to visual object recognition, which might be relevant to the code.

[7] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. - Introduces the U-net architecture used in biomedical image segmentation, which might be related to the code.

[8] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. - Discusses the ImageNet Large Scale Visual Recognition Challenge, which could be related to the dataset used in the code.

[9] Shin, H., Roth, H. R., Gao, M., Lu, L., Xu, Z., Nogues, I., ... & Summers, R. M. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. - Discusses the application of deep convolutional neural networks for computer-aided detection, which might be relevant to the code.

[10] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. - Introduces GoogLeNet, a deep convolutional neural network architecture, which might be related to the code.