USING IMAGEBERT TO IMPROVE PERFORMANCE OF MULTI-CLASS CHEST X-RAY CLASSIFICATION

A PREPRINT

Saptarshi Purkayastha
Department of BioHealth Informatics
IUPUI
Indianapolis, IN 46202
saptpurk@iupui.edu

Ananth R. Bhimireddy
School of Informatics and Computing
IUPUI
Indianapolis, IN 46202
anbhimi@iu.edu

Priyanshu Sinha
Mentor Graphics Pvt. Ltd.
priyanshu.sinha@outlook.com

Judy W. Gichoya
Department of Radiology & Imaging Sciences
Emory University
Atlanta, GA 30322
judywawira@emory.edu

July 2, 2020

ABSTRACT

Pulmonary edema is a medical condition which is often related to life-threatening heart-related complications. Several recent studies have demonstrated that machine learning models using deep learning (DL) methods are able to identify anomalies on chest X-rays (CXR) as well as trained radiologists. Yet, there are limited/no studies that have integrated these models in clinical radiology workflows. The objective of this project is to identify state-of-the-art DL algorithms and integrate the classification results into the radiology workflow, more specifically in a DICOM Viewer, so that radiologists can use it as a clinical decision support. Our proof-of-concept (POC) is to detect the presence/absence of edema in chest radiographs obtained from the CheXpert dataset. We implemented the state-of-the-art deep learning methods for image classification -ResNet50, VGG16 and Inception v4 using the FastAI library and PyTorch on 77,408 CXR which have classified the presence/absence of edema in the images with an accuracy of 65%, 70% and 65% respectively on a test dataset of about 202 images. The CXR were converted to DICOM format using the img2dcm utility of DICOM ToolKit (DCMTK), and later uploaded tothe Orthanc PACS, which was connected to the OHIF Viewer. This is the first study that has integrated the machine learning outcomes into the clinical workflow in order to improve the decision-making process by implementing object detection and instance segmentation algorithms.

Keywords Neural networks · Chest radiographs · Pulmonary Edema · Computer vision · Deep Learning Algorithms · Convolutional Neural Networks · DICOM viewer

1 Introduction

Pulmonary edema is a medical condition caused by excess fluid in the lungs and is often characterized by shortness of breath and difficulty breathing and can be fatal if not treated quickly. Often associated with heart problems, however this fluid accumulation can also occur due to exposure to toxins, trauma, and high altitudes[1]. Diagnosis is dependent on clinical signs and symptoms and radiological images.Imaging is the fundamental role in diagnosing pulmonary edema. Chest radiography and Lung ultrasound are the most common diagnostic tools used. Chest radiography combines relative low cost with panoramic views that allows for exclusion of other pulmonary diseases that would fall under differential diagnosis and henceforth is much more preferable[2]. However in emergency conditions CXR are often

viewed at beside by portable machines, particularly in the diagnosis of acute cases. In these cases, a high inter-observer variability of bedside CXR reading limits the diagnostic usefulness of the methodology and complicates the differential diagnosis. Also the main limitation of CXR's that they are not sensitive enough to rule out differential diagnosis and in the presence of an abnormal pattern[3]. A recent review found a "real-time" error rate among radiologists in their day-to-day practices averages 3-5% [4]. Cognitive errors, which account for 20–40% of the total, occur when an abnormality is identified but the reporting radiologist fails to correctly understand or report its significance, i.e. misinterpretation. The more common perceptual error (60–80%) occurs when the radiologist fails to identify the abnormality in the first place, but it is recognised as having been visible in retrospect[5]. Prolonged attention to a specific area on a radiograph ("visual dwell") increases both false negative and false positive errors. Reducing the viewing time for CXRs to less than 4 seconds also increases the miss rate[6]. Applications of deep learning algorithms for an improved interpretation of chest radiographs can reduce fatigue-based diagnostic errors from radiologists. It can be very useful in areas where there is a lack of expert radiologists.

Automation of radiological interpretation of reports at different working levels as a radiologists has the potential to improve clinical workflow and decision support. Several studies have successfully applied deep learning algorithms on Chest X-rays to detect various clinical manifestations of lung disorders like pneumonia, pulmonary tuberculosis, pulmonary effusion, pneumothorax, lung cancer etc. Paras Lakhani et al (2017) applied convolutional neural networks like AlexNet and GoogLeNet for automated classification of pulmonary tuberculosis from chest radiographs[7]. They have achieved an AUC of 0.99.Pranav Rajpurkar et al (2017) developed an algorithm called "CheXNet", which is a 121-layer convolutional neural network trained on a publicly available chest X-ray dataset to detect pneumonia[8]. It was found from their study that CheXNet exceeds the performance of radiologists on the F1 metric.The "CheXNeXt" was applied to detect the presence of fourteen different pathologies from frontal-view chest X-rays[9]. CheXNeXt's average time to interpret the images in the validation dataset (420 images) was only 1.5 minutes compared to 240 minutes for radiologists to interpret the images. CheXNeXt achieved radiologist-level performance on 11 pathologies and did not achieve radiologist-level performance on 3 pathologies namely, cardiomegaly, emphysema, and hiatal hernia.

Image classification is the basis for computer-vision tasks such as localization, detection, and segmentation. CNNs are gaining popularity in the current scenarios for their high accuracy in image recognition and classification. Application of CNNs in the field of radiology helps the radiologists in making an accurate diagnosis from chest X-rays and thus, achieving enhanced patient healthcare. Deep learning methods offer several benefits to the radiologists as well as the patients. Application of deep learning models can help the radiologists by undertaking supporting tasks thus, reduces the radiologist workload. These models also help the healthcare professionals in detecting patients who require urgent therapeutic intervention[10]. The current study aims to build an image classifier that classifies the images for the presence of pulmonary edema by using a convolutional neural network(CNN)and address the computer-vision problem to aid in decision support by making these classified images available to view by the radiologists through the DICOM Web (OHIF)viewer. Hence, in this study, we implemented the state-of-the-art deep learning models to classify the images based on the presence or absence of a medical observation which in our case was edema. In addition to just classification, we obtained the probabilities of the presence or absence of the condition which would help clinicians prescribe the appropriate medication or suggest the most suitable therapeutic intervention to the patients. This feature when made available on the DICOM viewer, would thus improve the clinical care process as well as helps in early detection of patients at high-risk. It will also improve the clinical workflow and time management.

A common Clinical radiology workflow consists of the imaging machines and reports that can be viewed from an EMR (1) A Picture Archiving and Communication System (PACS), a medical report viewer which follows the Digital Imaging and Communications in Medicine (DICOM) protocol and an electronic medical record (EMR) system. Orthanc is an open-source, lightweight, powerful, standalone DICOM server that is commonly used in healthcare settings and for medical research. Orthanc is designed in Belgia with an aim to ease the data management and DICOM scripting for medical research and clinical routine[11]. Orthanc brings the DICOM images to the community of computer vision and eases the analysis of clinical reports. (2)DICOM is the acronym for Digital Imaging and Communications in Medicine. All the reports generated by various medical hardware modalities, like ultrasound, X-rays, computed tomography (CT) and 4magnetic resonance imaging (MRI) machines will be stored in a PACS system which in turn will be viewable through a DICOM viewer. The LibreHealth EMR system consists of a radiology module which follows the common radiology workflow (as discussed above) and is connected to Orthanc (PACS) and a DICOM Web (OHIF) viewer via a docker image. The following figure presents the LibreHealth radiology workflow.Fig 1: LibreHealth Radiology WorkflowThe viewer within the LibreHealth is based on cornerstone, which is a lightweight, browser-based JavaScriptlibrary for medical imaging using HTML5 canvas3. The viewer is connected to the PACS through DICOMWeb. The PACS system runs based on dcm4che and Orthanc. (3)dcm4che is a collection of open source applications and utilities for the healthcare enterprise. dcm4che is developed in the Java programming language for performance and portability supporting deployment in JDK. 1.6 and up(https://www.dcm4che.org/)4. DICOMWeb

and DICOM MPPS connects the PACS with the radiology imaging system (RIS). The RIS in LibreHealth consists of the radiology application programming interface (API) in the backend and a user interface(UI) which is anopen web app (OWA). The fast healthcare interoperability resources (FHIR) and Representational state transfer (REST) connect the RIS to the EMR which is a LibreHealth Toolkit.

With the wide range of applications on medical diagnosis is available by machine learning does have popularity, but deep learning has improved the performance of various machine learning tasks compared to the traditional machine learning algorithms like artificial neural networks. DL is interesting not only because its level of performance is greater, but also because it does not require a human to identify and compute the critical features. Instead, during training, DL algorithms "learn" discriminatory features that best predict the outcomes. The amount of human effort required to train DL systems is less as there is a requirement of feature engineering or computation. Deep learning may also lead to the discovery of important new features that were not anticipated[12]. Some forms of DL can accurately segment organs, essentially, trace the boundaries, enabling volume measurements or calculation of other properties. Deep Learning networks have the capacity to predict important properties from regions of an image. Though Deep Learning offers many advantages, it does have the underlying challenges. Training a deep learning algorithm requires more data and more care while analyzing the results. It is impossible to understand how a neural network works to arrive at a solution. This “lack of transparency” makes it hard to predict when failures might occur. For instance, many studies have proven that deep learning algorithms can predict physical anomalies of the human body better than doctors. But the machine must provide justification and reasoning for its prediction in order to reassure its accuracy. It would be difficult to gain the trust of patients or learn why any mistakes in diagnosis were made without such justification/reasoning. Overfitting is another challenge with neural networks. Currently, deep learning techniques are state-of-the-art for classification of images. Deep convolutional neural network (DCNN) is a type of deep learning approach which is well suitable for image analysis. Interest in DCNNs for image classification has begun to raise from 2012 through ImageNet Large Scale Visual Recognition Challenge[13].This resulted in a decrease in the classification error rate from approximately 25% in 2011 to 3.6% in 2015[14].

2 Methodology

Overview

In this section we discuss about the mentions about the tools, packages/software, details of gathering the data, data description, extraction, image pre-processing, and implementation of the models.

2.1 Data source and Description

For this study, we used the CheXpert dataset obtained from the Stanford Machine Learning group. CheXpert is the largest dataset of chest X-rays and competition for automated interpretation of chest x-rays. This dataset consists of a total of 224,316 chest radiographs of about 65,240 patients. The data was collected from Stanford Hospital over about fifteen years i.e., from October 2002 to July 2017. It features uncertainty labels and radiologist-labeled reference standard evaluation sets15. Each radiology report was labelled for the presence of 14 observations as positive, negative or uncertain. The fourteen observations were no finding, enlarged cardio mediastinum, cardiomegaly, lung lesion, lung opacity, edema, consolidation, pneumonia, atelectasis, pneumothorax, pleural effusion, pleural other, fracture and support devices. The final label for each observation is blank for unmentioned, 0 for negative, 1 for positive and -1 for uncertain. The dataset consisted of a total of 19 columns, among which 14 columns has the information related to the fourteen different medical observations as mentioned above and the rest of the columns were path, sex, age, frontal/lateral and anteroposterior (AP)/posteroanterior (PA). The “path” column provides information about the complete path of the image within the folders in the dataset for easily finding the location of the image while the other two columns by name, “frontal/lateral” and “AP/PA” mentions the different views of the radiograph and the position of the film plate & the X-ray machine respectively.

2.2 Data Extraction and pre-processing

The downloaded data consisted of training data (train.csv), its associated images, validated data (valid.csv) and its associated images. We did data analysis on the dataset that the researchers have used to train their model (i.e., train.csv) to identify the top five observations using the package “pandas” in python. Firstly, the data were separated based on the gender of the patients. Out of 223,414 reports, there were 90,777 females, 132,636 males and one unknown gender. We excluded the patients of unknown gender and the patients less than 18 years of age. The top five features containing either 1,0 or -1 values as labels were pleural effusion (133,212), support devices (123,218), lung opacity (117,779), edema (85,956), and pneumothorax (78,935).Among the images that have edema labels, 52,246 images contained

positive labels that indicate the presence of edema. We labeled 20,726 images has "0" as a label which indicates the absence of edema and the rest 12,984 images as "-1," indicating the uncertainty of edema. Inclusion Criteria: The final data that we considered for this project included male and female patients with the label “edema” in the age range of 18 years to 60 and greater than 60 years. Furthermore, we have considered only frontal view images irrespective of the film position, i.e., AP/PA. A total of 77,408 images met the inclusion and exclusion criteria.

Out of the 77,408 images, 49,675 images have positive labels, 15,915 images have negative labels, and the rest 11,818 images have uncertainty labels for edema. Similarly, the validation images were filtered to contain frontal view images with all three edema labels. There were 202 validation images out of 234 images that met the study criteria. Out of these 202 images, 160 images have "0" as the label (negative), while the remaining 42 images have "1" as the label (positive).

2.3 Image Pre-processing

The images presented by the CheXpert dataset were in.JPG format. In order to view them in OHIF (Open Health Imaging Foundation) viewer and Orthanc (Picture Archiving and Communication System, PACS), these images must be in DICOM format.DCM extension which was achieved using the img2dcm utility of DICOM ToolKit (DCMTK) in the terminal of Jupyter notebook. Once the images were successfully converted to.DCM, we added few required metadata elements like patient name, gender, age, modality and study description according to the elements defined in the DICOM dictionary using the pydicom module in python. As the original dataset did not have any patient names, we have randomly assigned a list of ten names i.e., five female (Kelsey, Mary, Lisa, Erica and Olivia) and five male names (Sunny, James, John, George and Harry) to the DICOM images for a better representation in the OHIF viewer.

2.4 Model Implementation

We used Fastai (version 1.4.1) which is a new open source deep learning library introduced by fast.ai. Fastai library is built on the top of PyTorch which a scientific computing package based on Python to implement convolutional neural network models for image classification (Jeremy Howard, 2018)23. PyTorch is one of the preferred deep learning research platforms that make use of the power of Graphic Processing Units (GPUs) to build the neural network models effortlessly24. We trained and tested the models on NVIDIA’s graphic processing unit. This study has chosen to implement ResNet50, VGG16 and Inception4 models for image classification. There are more sophisticated and better performing deep learning models like Faster Region Convolutional Neural Network (Faster RCNN), You Only Look Once (YOLO) and Mask Region Convolutional Neural Network (Mask RCNN) which are specifically meant for object detection and instance segmentation. But our data was not suitable to run those models as it has no bounding box annotations and image segmentation.

2.4.1 ResNet50 model

We created a image classifier using the convolutional neural network (CNN) in order to classify the images based on their respective labels (1,0, -1) using ResNet50 (Residual Network 50) architecture. Residual Network is a powerful backbone model and it is often used for performing many computer visions tasks. ResNet model was the winner of 2015’s ImageNet challenge. ResNet allows training the deep neural networks with over 150 layers. Before ResNet, it was hard to train deep networks because of the vanishing gradient problem. With the increasing depth of the network, the performance of the model starts degrading or gets saturated. ResNet models have solved this issue as they make use of the concept of skip connection24. Firstly, the dataset of a total of 77,408 chest radiographs which is available as a .csv file was imported into the Jupiter notebook. Then, the numerical labels 1,0 and -1 for the presence, absence and uncertainty of edema were replaced with the corresponding terms namely, edema, no edema and uncertain. The images were clipped to the size 128 for training the model.

The dataset was split into training and validation sets containing 80% and 20% data respectively. The ResNet50 model was trained on 80% of the training dataset (obtained from the original dataset) which consisted of about 61,927 images while the validation dataset consisted of the rest 15,481 images. Initially, the model was trained for five epochs at a learning rate of 0.0001. The metrics calculated were training loss, validation loss, accuracy and the time taken to complete each epoch. At the end of five epochs, the training loss was 78%, validation loss was 76% and the accuracy of classification was 69.6%. The model was trained for ten more epochs at the same learning rate of 0.0001. Here, the accuracy of classification was improved (73.0%) with the training loss of 74.4% and the validation loss of 73.8%. Likewise, the model was trained for 20 more epochs at the same learning rate and reported an accuracy of 71.2%. But the validation loss (71.8%) exceeded the training loss (70.7%). Then, the model was unfrozen and a learning rate of 0.00001 was used to train the model for 10 epochs. Here, the training loss was found to be greater than the validation loss with the accuracy of 71.3%. The model was thus trained for 45 epochs and the total time taken to run all the 45

epochs was 78 minutes 22 seconds . The number of epochs, training loss, validation loss and accuracy at the end of each set of epochs are presented in Table 1.

Once the model was trained, we introduced the test dataset consisting of 202 chest x-ray images to test the efficiency of the model in classification of the presence or absence of edema on test images. The images were first clipped to the size of 128. The function, "ClassificationInterpretation" was used to obtain the probabilities of images for the presence of edema and the function.

2.4.2 VGG16 model

The second model we implemented was VGG16, one of the famous models submitted to ILSVRC-2014. VGG16 is a CNN model introduced by K. Simonyan and A. Zisserman from the University of Oxford in the paper "Very Deep Convolutional Networks for Large-Scale Image Recognition". The model is pre-trained on ImageNet data and it has achieved an accuracy of 92.7% top-5 test accuracy in ImageNet. VGG16 makes use of NVIDIA Titan Black GPU's25.

The dataset consisting of a total of 77,408 images which is available as a .csv file was imported into the Jupiter notebook using pandas. Then, the numerical labels 1,0 and -1 for the presence, absence and uncertainty of edema were replaced with the corresponding terms namely, edema, no edema and uncertain. The images were clipped to the size 128 for training the model. The dataset was split into training and validation sets containing 80% and 20% data respectively. The VGG16 model was trained on 80% of the training dataset (obtained from the original dataset) which consisted of about 61,927 images while the validation dataset consisted of the rest 15,481 images. Initially, the model was trained for ten epochs at a learning rate of 0.0001. The metrics calculated were training loss, validation loss, accuracy and the time taken to complete each epoch. At the end of the first ten epochs, the training loss was 77.8%, validation loss was 74.7% and the accuracy of classification was 70.4%. The model was trained for ten more epochs at the same learning rate of 0.0001. At this stage, the training loss and validation loss were found to be almost the same and the accuracy was found to be 70.8%. Likewise, the model was trained for 30 more epochs until the training loss is greater than the validation loss. The model was trained for a total of 50 epochs and the total time taken to run all the 50 epochs was 121 minutes 26 seconds.

Once the model was trained,we introduced the test dataset consisting of 202 chest x-ray images to test the efficiency of the model in classification of the presence of edema on test images. The images were first clipped to the size of 128. Then, the function "ClassificationInterpretation" was used to obtain the probabilities of images for the presence of edema.Both ResNet50 and VGG16 were implemented in Fastai version 1.4.1.

2.4.3 Inception v4 model

: We augumented the training data using random resized image cropping and horizontal flip whereas validation set was augmented using center cropping and normal resizing technique. We used state-of-the-art inception v4 model for image recognition implemented by Google. Inception v4 is a simplified version of its previous inception (i.e., v2, v3) networks. In Inception v4, the "stem" block is modified which performs initial set of operations before passing to the three main inception modules (A, B, C) and reduction blocks have been introduced which change height and width of the grid[15]. Thus, the inception modules (A, B, C) in v4 were made more uniform than their previous complicated versions to boost the performance. We have used Cadene's pytorch inception v4 pretrained model which was trained on the massive ImageNet data30. The last layer of the pretrained model was modified to output only 3 classes. Our model uses stochastic gradient descent optimizer (SGD) and binary cross entropy loss function to calculate the error rate between predicted and true values. We have trained our model for 15 epochs with a batch size of 8 and learning rate of 0.001. Our model has achieved 73.51 % accuracy for the validation set. There was no overfitting problem as the training accuracy and validation accuracy were almost similar i.e., 73.04 and 73.51 respectively. The total time taken for training all the 15 epochs was 472 minutes 52 seconds.

Once the model was trained,we tested it on the test dataset consisting of 202 chest x-ray images to test the efficiency of the model in classification of the presence/absence of edema on test images. The images were first clipped to the size of 128. The last layer of the model was changed to SoftMax instead of cross entropy loss. The function "ClassificationInterpretation" was used to obtain the probabilities of images for the presence of edema.

2.5 Librehealth Radiology System and OHIF viewer

2.5.1 Librehealth Radiology System

Firstly, we downloaded and deployed lh-radiology module on a local machine. There are many options to install the LibreHealth radiology RIS system. One of the easiest ways is to use Docker. The other way is to build with Maven and Java JDK8. About Docker Docker is an open source tool designed to make it easier to create, deploy, and run

applications by using containers which allow a developer to package up an application with all the parts it needs, such as libraries and other dependencies, and ship it all out as one package33. Docker Installation First, download and install Docker Desktop from docker hub33,34. It is available for Windows as well as MacOS. Once the Docker Desktop is downloaded, double-click "Docker for Windows Installer" to run the installer. When the installation finishes, Docker starts automatically. The whale in the notification area indicates that Docker is running, and accessible from a terminal. Docker is available in any terminal as long as the Docker Desktop - Windows app is running. To build the radiology module and its Docker image, docker-compose need to be installed by executing the following code in the command prompt.

Once the Docker-compose is successfully installed, Librehealth Radiology will be accessible at http://localhost:8080/lh-toolkit/.

2.6 Building and deploying the OHIF viewer Meteor App

To build and deploy the OHIF Viewer,we installed Meteor[16].Meteor is an open source platform for web, mobile and desktop in pure JavaScript. Meteor allows to ship more with less code as it has an integrated JavaScript stack that extends from the database to the end user's screen. Meteor builds apps for any device (iOS, Android, desktop etc.,) using the same code. It integrates the technologies which are already in use and focuses on building features instead of requiring the users to wire disparate components together36. After installation of the meteor, the following we did the following: Cd Viewers/OHIFViewer npm install Linux: ./bin/orthancDICOMWeb.sh OR Windows: ./bin/orthancDICOMWeb.bat

These steps will deploy OHIF Viewer on http://localhost:3000 and will be available from the Radiology header menu of the LibreHealth Toolkit.

2.7 Addition of an extra dicom tag to the image viewport

As mentioned earlier, the prime focus of our project is to show the probabilities of the edema in the OHIF viewer. In order to do so, we need to add an extra DICOM tag called "image comments" to the image viewport on OHIF. The libre health radiology folder that was downloaded from GitHub has a folder called "Viewers" within which there is a folder by name "packages". Then we directed to the following folders in the order, ohif-cornerstone, client, lib, classes, and finally the JSON file "MetadataProvider.js" to which the tag "image comments" with the value 0020,4000 was added under the element, imageMetadata.instance.imageComments. Then we added the tag and its value to the JSON file by name "retrieveMetadata.js" which is contained in the folders by the order, Viewers, packages, ohif-studies, imports, server, services, and remote. We had to add the tag to the "viewportOverlay.html" and "viewportOverlay.js" which are contained within the folders in the order, Viewers, Packages, ohif-viewerbase, client, components, viewer, and viewportOverlay.

3 Results

S.No.	Model	Accuracy(%)	Precision	Recall	F-1 Score	AUROC
1	ResNet50	65	0.37	0.95	0.53	0.76
2	VGG16	70	0.40	0.86	0.55	0.74
3	Inception V4	65	0.37	0.95	0.53	0.76

Table 1: Performance of different models

ResNet50 has classified the test images with an accuracy of 65%. The precision and recall for the presence of edema were 0.37 and 0.95 while the precision and recall for the absence of edema were 0.98 and 0.57 respectively. The F1 score for the presence of edema was 0.53 and 0.72 for the absence of edema. The AUROC was found to be 0.76. Similar results were observed for inception v4. The model has classified the images with an accuracy of 65%. The precision and recall for the presence of edema were 0.37 and 0.99 while the precision and recall for the absence of edema were 0.95 and 0.57 respectively. The F1 scores for the presence and absence of edema were 0.53 and 0.73 respectively. The AUROC was found to be 0.76.VGG16 has classified the images with 70% accuracy. The precision for the presence of edema was 0.40 and 0.97 for the absence of edema. The recall for the presence of edema was 0.86 while 0.66 for the absence of edema. The F1 scores for the presence and absence of edema were 0.55 and 0.79 respectively. The AUROC was found to be 0.74.

We have evaluated the performance of the models based on their AUROC values. Though VGG16 has classified the test images with more accuracy compared to ResNet50 and Inception v4, it has a low AUROC value which is why it is not a good classifier for our dataset. ResNet50 and Inception v4 gave better results compared to VGG16. We have

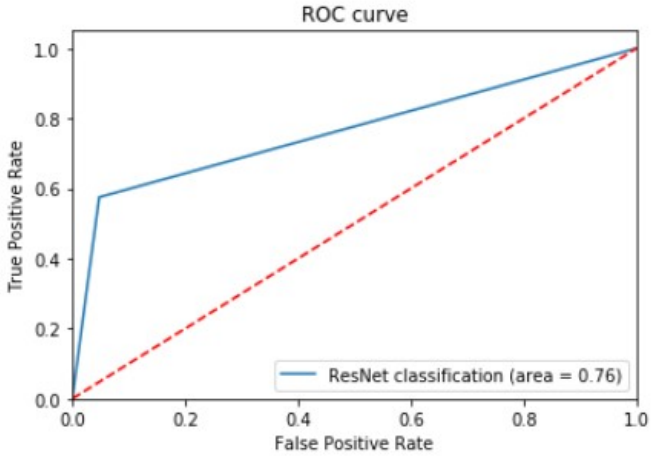

Figure 1: ROC curve of ResNet50

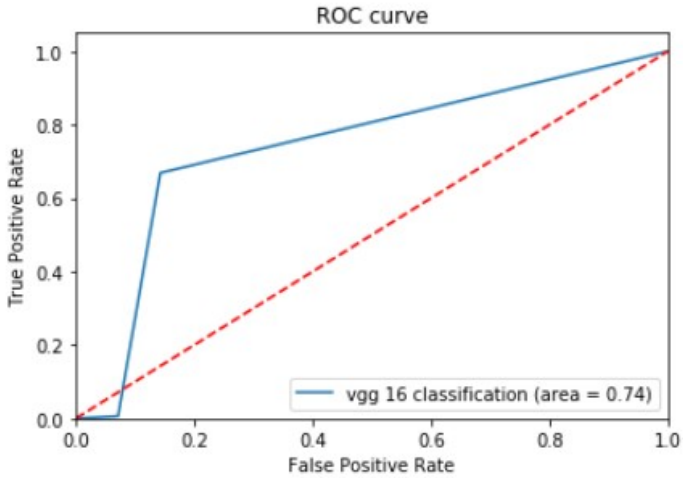

Figure 2: ROC curve of VGG16 classification

chosen to upload the predictions made by ResNet50 to the Orthanc instead of Inception v4 as the size of the architecture, complexity and time taken to run ResNet50 were much less compared to Inception v4.

We used the in-built method, "get_preds" to obtain the probability predictions for each image in the test dataset. The probability predictions came out in the form of a tensor with three values for each image i.e., for edema, no edema and uncertain. Then the highest probability out of the three values for each image was filtered out and tagged with appropriate label. The image names along with their probabilities and tags were written to a csv file . The test images which were converted earlier to .dcm format using img2dcm were stored in a folder, “edited_valid_dcm”. The image names in the csv file were matched with the file names in the DICOM folder in order to get the probability as well as the respective tag into the edited_valid_dcm folder. The new tag “image comments” was added to reflect the value of probabilities on OHIF viewer using the PyDicom module of python. The files in the dicom folder were converted to a zip file. All the images contained in the zip file were uploaded to Orthanc using “upload” option in the top-right corner on the home screen of Orthanc. After a successful upload, these images reflect in the “lookup” section of Orthanc as well as under the “study list” of OHIF viewer. We can access the images through the LibreHealth Radiology module all of which run through a docker. The output that is visible on OHIF viewer is presented in the following figures. For instance, in Fig 15, the probability of edema is 0.85 while in Fig, the probability of no edema is 0.76.

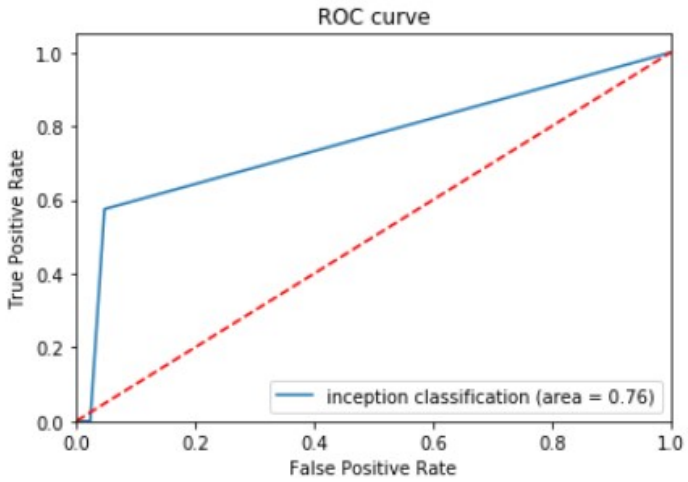

Figure 3: ROC curve of Inception v4 Model

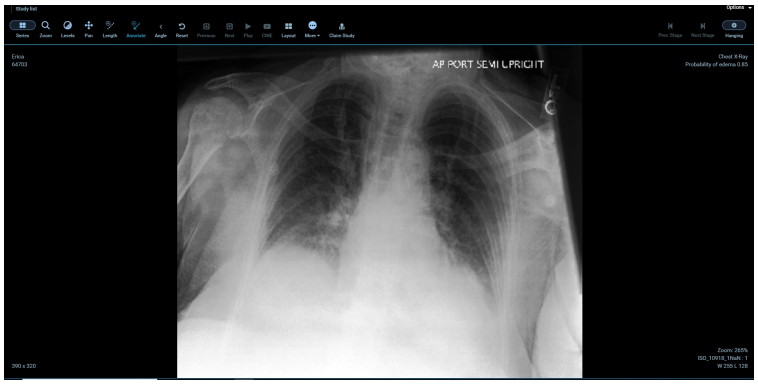

Figure 4: Image showing the predicted probability of edema in OHIF Viewer

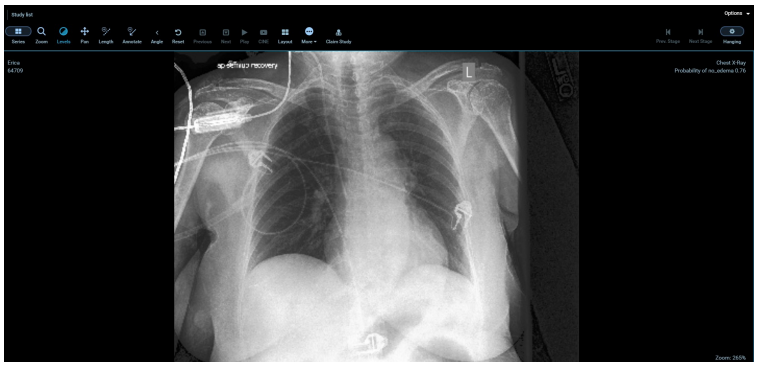

Figure 5: Image showing the predicted probability of no_edema in OHIF Viewer

4 Discussion

Pulmonary edema is the underlying cause for several fatalities. It is often possible to detect edema from chest X-rays. Interpretation of a chest X-ray is a tedious and complex task for doctors. Delay in detection of edema could lead to further delays in clinical diagnosis and treatment thus, jeopardizing the life of the patients. Development of automated detecting systems using deep learning methods helps in faster detection of the anomalies with greater accuracy to aid human readers and this would be beneficial at times where there is no adequate staffing in the healthcare center.

Automated systems also help in prioritizing the patients who need an immediate therapeutic intervention. Deep convolutional neural networks are outperforming humans in image classification and visual recognition tasks. As our project chose to deal with providing a solution to computer vision recognition task, deep convolutional neural networks were selected attributed to their efficiency and good performance in image classification. For our study, ResNet50 which has 50 layers has given a classification accuracy of 65% and an F1 score of 0.53 for the presence of edema and 0.72 for the absence of edema in the test images. The F1 score for detecting the absence of edema is higher in the validation set and this might be attributed to the imbalance in our dataset i.e., our dataset has a greater number of no edema images compared to edema images. Similar is the case with the other two models. A Chinese prospective, randomized study has implemented deep learning methods to build a system for automatic detection of polyps in the real clinical setting. The authors have used SegNet architecture to which the input is a colonoscopy image while the output is a hollow tracing box on the CADe monitor representing the probability of a polyp which is much similar to our study except the incorporation of tracing boxes[17]. Deep learning and machine learning approaches are being widely used to detect the diseases from chest X-rays. Andrew G. Taylor et al in 2018 implemented deep learning models like VGG16, VGG19, Inception, Xception and ResNet to automate the detection of pneumothorax in frontal chest X-rays. This study has achieved the AUCs of 0.94 and above for correct detection of moderate and large pneumothoraces from chest radiographs[10]. A study has attempted to detect twelve common pathologies (among which edema is one) that are possible in the chest X-rays by implementing backpropagation neural network, competitive neural network and convolutional neural network. CNN has achieved a highest recognition rate of 92.4% compared to the other networks which is attributed to its deep structure. Also, CNNs outperforms the models based on transfer learning like VGG16 and VGG19 which was also evident in out study[18]. Islam et al, 2017, studied the performance of deep convolutional networks like ResNet, AlexNet and VGG in automatic detection and localization of about 20 abnormalities in chest X-rays using three publicly available chest X-ray datasets[19]. It is evident from this study that ResNet50 gives high accuracy and AUC in classifying pulmonary edema compared to other models. They also employed ensemble methods to achieve higher classification accuracy To our knowledge, this study is the first study to integrate the outcomes of deep learning algorithms with a DICOM viewer in order to aid clinicians in automatic detection of edema in the chest X-rays. The results represent the respective probabilities of edema and no edema under image comments section on OHIF viewer. Ensemble of different algorithms could improve the classification accuracy as well as the AUROC. Future studies can improve the image classification by incorporating bounding box annotations with the help of a radiologist for a better clinical decision support tool.

5 Conclusion

In this study, we have observed the performance of three models on the CheXpert dataset in classifying the images that have edema. From our project, it is evident that ResNet50 and Inception4 have performed similarly on our test dataset. The accuracy of ResNet50 and Inception4 on test dataset was 65% while VGG 16 reported an accuracy of 70%. But the ROC values of ResNet50, VGG16 and Inception4 were 0.76, 0.74 and 0.76 respectively. In addition to just image classification by using deep learning models, our project has showed the probability of the feature under study i.e., edema on OHIF viewer. We hope that our implementation of OHIF viewer will serve the radiologists in making appropriate clinical decisions and this would also serve as a basis for further research in improving the decision-making process by implementing several other novel models for object detection and instance segmentation.

5.1 Future Directions

Future work can focus on creating bounding box annotations with the help of radiologists for the images contained in the dataset. This makes it possible to run more sophisticated and faster models with special emphasis on image segmentation, object detection, and instance segmentation. In addition, future researches can focus on showing all the bounding boxes in the OHIF viewer and make them editable by the radiologists if possible, for an even better clinical decision support system. Future researches can even focus on building a better image classifier by ensembling the models for good accuracy, ROC and F1 scores on a huge dataset and make the probabilities viewable on the OHIF viewer. Future researches can make use of the Orthanc REST API to upload the images and the predictions to Orthanc.

References

[1] Chest X-rays - Mayo Clinic.

[2] Ana Carolina Peçanha Antonio, Cassiano Teixeira, Priscylla Souza Castro, Ana Paula Zanardo, Marcelo Basso Gazzana, and Marli Knorst. Usefulness of radiological signs of pulmonary congestion in predicting failed spontaneous breathing trials. *J Bras Pneumol*, 43(4):253–258, 2017.

[3] Luciano Cardinale, Giovanni Volpicelli, Alessandro Lamorte, and Jessica Martino. Revisiting signs, strengths and weaknesses of Standard Chest Radiography in patients of Acute Dyspnea in the Emergency Department. *J Thorac Dis*, 4(4):398–407, August 2012.

[4] Adrian Brady, Risteárd Ó Laoide, Peter McCarthy, and Ronan McDermott. Discrepancy and Error in Radiology: Concepts, Causes and Consequences. *Ulster Med J*, 81(1):3–9, January 2012.

[5] Michael A. Bruno, Eric A. Walker, and Hani H. Abujudeh. Understanding and Confronting Our Mistakes: The Epidemiology of Error in Radiology and Strategies for Error Reduction. *Radiographics*, 35(6):1668–1676, October 2015.

[6] P. J. Robinson. Radiology's Achilles' heel: error and variation in the interpretation of the Röntgen image. *Br J Radiol*, 70(839):1085–1098, November 1997.

[7] Paras Lakhani and Baskaran Sundaram. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. *Radiology*, 284(2):574–582, April 2017.

[8] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv:1711.05225 [cs, stat]*, December 2017. arXiv: 1711.05225.

[9] Pranav Rajpurkar, Jeremy Irvin, Robyn L. Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P. Langlotz, Bhavik N. Patel, Kristen W. Yeom, Katie Shpanskaya, Francis G. Blankenberg, Jayne Seekins, Timothy J. Amrhein, David A. Mong, Safwan S. Halabi, Evan J. Zucker, Andrew Y. Ng, and Matthew P. Lungren. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLOS Medicine*, 15(11):e1002686, November 2018.

[10] Andrew G. Taylor, Clinton Mielke, and John Mongan. Automated detection of moderate and large pneumothorax on frontal chest X-rays using deep convolutional neural networks: A retrospective study. *PLOS Medicine*, 15(11):e1002697, November 2018.

[11] Sébastien Jodogne. The Orthanc Ecosystem for Medical Imaging. *J Digit Imaging*, 31(3):341–352, June 2018.

[12] Bradley J Erickson, Panagiotis Korfiatis, Timothy L Kline, Zeynettin Akkus, Kenneth Philbrick, and Alexander D Weston. Deep learning in radiology: does one size fit all? *Journal of the American College of Radiology*, 15(3):521–526, 2018.

[13] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9):2352–2449, 2017.

[14] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[15] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. *arXiv:1602.07261 [cs]*, August 2016. arXiv: 1602.07261.

[16] Meteor Development Group. Build Apps with JavaScript | Meteor.

[17] Pu Wang, Tyler M. Berzin, Jeremy Romek Glissen Brown, Shishira Bharadwaj, Aymeric Becq, Xun Xiao, Peixi Liu, Liangping Li, Yan Song, Di Zhang, Yi Li, Guangre Xu, Mengtian Tu, and Xiaogang Liu. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut*, 68(10):1813–1819, 2019.

[18] Rahib H. Abiyev and Mohammad Khaleel Sallam Ma'aitah. Deep Convolutional Neural Networks for Chest Diseases Detection, 2018.

[19] Mohammad Tariqul Islam, Md Abdul Aowal, Ahmed Tahseen Minhaz, and Khalid Ashraf. Abnormality Detection and Localization in Chest X-Rays using Deep Convolutional Neural Networks. *arXiv:1705.09850 [cs]*, September 2017. arXiv: 1705.09850.