Application and Evaluation of Deep Learning for Detecting Lung Cancer Tumors using Crowdsourcing

Undergraduate Thesis

Submitted in partial fulfillment of the requirements of
BITS F421T Thesis

By

Aakanksha Sanctis
ID No. 2015B3A70530G

Under the supervision of:

Dr. Amrapali Zaveri, Dr. Deniz Iren
&
Dr. Neena Goveas

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI, GOA CAMPUS
December 2019

Declaration of Authorship

I, Aakanksha SANCTIS, declare that this Undergraduate Thesis titled, 'Application and Evaluation of Deep Learning for Detecting Lung Cancer Tumors using Crowdsourcing ' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: 

Date: 06/12/2019

Certificate

This is to certify that the thesis entitled, "*Application and Evaluation of Deep Learning for Detecting Lung Cancer Tumors using Crowdsourcing* " and submitted by Aakanksha SANCTIS ID No. 2015B3A70530G in partial fulfillment of the requirements of BITS F421T Thesis embodies the work done by her under my supervision.

Supervisor
Dr. Amrapali ZAVERI
Post Doctoral Researcher,
Institute of Data Science, Maastricht University, Netherlands
Date:

On-campus Supervisor
Dr. Neena GOVEAS
Professor,
BITS Pilani, K.K Birla Goa Campus
Date:

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI, GOA CAMPUS

Abstract

B.E (Hons.) Computer Science, MSc. (Hons.) Economics

Application and Evaluation of Deep Learning for Detecting Lung Cancer Tumors using Crowdsourcing

by Aakanksha Sanctis

In this research, we present a crowdsourcing approach for detecting Lung cancer tumors using deep learning. With the advances in deep learning, many Computer Assisted Diagnostic (CAD) systems have been developed but the number of annotated images for training such deep learning models is limited compared to natural image datasets. Medical image annotation is quite expensive, as it requires the time of experienced radiologists and doctors. To address this challenge, we propose a complete workflow for a refined CAD, which includes the collective intelligence of the crowd (non-experts) to annotate the huge amount of medical data specifically Lung CT scans and evaluate the feasibility of the crowd for such tasks. To analyze the best design parameters to accurately annotate lung nodules in CT scan images, we conduct four different experiments with the crowd. To check how 'experts' fare on the same experiment as compared to 'non-expert' crowd, we repeat the same experiment with experts as well. Annotations from these experiments are aggregated using DBSCAN clustering and the crowd agreement is compared with expert annotations. Aggregation of crowd annotations produced cluster centroids which had high correlation as compared to expert annotations.The proposed workflow have been evaluated on LUNA16 dataset. The crowd annotated LUNA16 dataset will be used to train a deep learning model and results will be compared to evaluate if a crowdenabled deep learning model can perform better than state of the art deep learning models.

Acknowledgements

I would like to thank my supervisors Dr. Amrapali Zaveri and Dr. Deniz Iren for guiding me and giving me the opportunity to collaborate on this significant project. I would like to thank Prof Michel Dumontier for providing his valuable suggestions on how to approach this complex problem. I would like to thank the researchers at D-Lab, Maastricht University, for providing insightful inputs to approach a medical diagnosis solution. I am grateful to all my colleagues at Institute of Data Science (IDS), Maastricht University, for supporting me throughout my thesis. Last but not the least, I would like to thank my on-campus supervisor Prof Neena Goveas and my institute BITS Goa for guiding me and molding me throughout my undergraduation journey.

Contents

Declaration of Authorship i

Certificate ii

Abstract iii

Acknowledgements iv

Contents v

List of Figures vii

List of Tables viii

Abbreviations ix

1 Introduction 1

2 Related Work 3
- 2.1 Crowdsourcing . . . 3
- 2.2 Automated tumor detection systems . . . 4
  - 2.2.1 Image Processing based systems . . . 4
  - 2.2.2 Deep Learning based systems . . . 4
- 2.3 Active Learning . . . 5

3 Methodology 6
- 3.1 Dataset . . . 6
- 3.2 Crowdsourcing . . . 7
  - 3.2.1 Preprocessing for crowdsourcing experiments . . . 7
  - 3.2.2 Experiments . . . 9
    - 3.2.2.1 Sample selection . . . 10
    - 3.2.2.2 Experimental design for crowdsourcing . . . 11

4 Results and Discussion 13
- 4.1 Crowdsourcing . . . 13
  - 4.1.1 Worker wise performance for crowdsourcing tasks . . . 14

4.1.1.1 Experiment 1: Unsegmented sequence subtasks . . . 14
4.1.1.2 Experiment 2: Segmented sequence subtasks . . . 15
4.1.1.3 Experiment 3: Segmented random subtasks . . . 16
4.1.1.4 Experiment 4: Unsegmented random subtasks . . . 17
4.1.1.5 Experiment with experts . . . 18
4.1.2 Aggregation of crowdsourced annotations per image . . . 20
4.1.2.1 Comparison between aggregated point and ground truth . . . 20

5 Conclusion **23**
5.1 Limitations . . . 23
5.2 Future work . . . 24
5.2.1 Deep Learning based automatic detection system . . . 24
5.2.1.1 Model Selection . . . 25
5.2.2 Active Learning . . . 25

A Appendix **27**
A.1 Experiments for crowdsourcing . . . 27
A.1.1 Unsegmented sequence experiment . . . 27
A.1.2 Segmented Sequence experiment . . . 27
A.1.3 Unsegmented random experiment . . . 27
A.1.4 Segmented random experiment . . . 27
A.2 Examples of aggregation of annotations . . . 29
A.2.1 Negative sample image with no clusters (TN) . . . 29
A.2.2 Negative sample image with clusters (FP) . . . 29
A.2.3 Positive sample image with cluster exactly matching true nodule (TP) . . . 30
A.2.4 Positive sample images with both TP and FP . . . 30
A.2.5 Positive sample image with no cluster (FN) . . . 31
A.3 Aggregation results for crowdsourced experiments using DBSCAN clustering algorithm . . . 32
A.3.1 Unsegmented sequence experiment . . . 32
A.3.2 Segmented sequence experiment . . . 33
A.3.3 Segmented random experiment . . . 34
A.3.4 Segmented sequence experiment with experts . . . 35

Bibliography **36**

List of Figures

1.1 Proposed architecture of the project . . . 2

3.1 Detailed pipeline of crowdsourcing experiments . . . 7
3.2 Preprocessing steps of 2D slice of lung CT scan . . . 8
3.3 Correlation heatmap of nodule characteristics . . . 9
3.4 Task presentation in the custom application for annotating lung tumors . . . 10
3.5 Parameters for crowdsourcing experiments . . . 12

4.1 Annotation parameters for an image . . . 13
4.2 Worker performance for unsegmented sequence experiment . . . 15
4.3 Worker performance for unsegmented sequence experiment with only positive samples . . . 16
4.4 Worker performance for segmented sequence experiment with only positive samples 16
4.5 Worker performance for segmented random experiment with only positive samples 17
4.6 Worker performance for unsegmented random experiment with only positive samples 17
4.7 Worker performance with experts for segmented sequence experiment with only positive samples . . . 18
4.8 Crowd annotations on a CT slice . . . 20
4.9 Crowd agreement on negative sample . . . 21
4.10 Crowd agreement on positive sample . . . 21
4.11 Aggregated annotation results for all crowdsourcing experiments . . . 22

5.1 Dual Path connections . . . 24

A.1 Unsegmented sequence experiment . . . 27
A.2 Segmented sequence experiment . . . 28
A.3 Unsegmented random experiment . . . 28
A.4 Segmented random experiment . . . 28
A.5 Negative sample image with no clusters (TN) . . . 29
A.6 Negative sample image with clusters (FP) . . . 30
A.7 Positive sample image with cluster exactly matching true nodule (TP) . . . 30
A.8 Positive images with both TP and FP . . . 31
A.9 Positive sample image with no cluster (FN) . . . 31
A.10 Aggregation results for Unsegmented sequence experiment . . . 32
A.11 Aggregation results for segmented sequence experiment . . . 33
A.12 Aggregation results for segmented random experiment . . . 34
A.13 Aggregation results for segmented sequence experiment using experts . . . 35

List of Tables

3.1 Comparison of types of annotation 11

Abbreviations

LUNA	**Lu**ng **N**odule **A**nalysis
LIDC-IDRI	**L**ung **I**mage **D**atabase **C**onsortium image collection
NLST	**N**ational **L**ung **S**creening **T**rial
CAD	**C**omputer **A**ided **D**iagnosis
CT	**C**omputerized **T**omography
CNN	**C**onvolutional **N**eural **N**etwork
HIT	**H**uman **I**ntelligence **T**ask
KW	**K**nowledge **W**orker
FP	**F**alse **P**ositive
TS-MIPS	**T**hin **S**lab **M**aximum **I**ntensity **P**rojections
ROI	**R**egion **o**f **I**nterest
FROC	**F**ree **R**esponse **O**perating **C**haracteristics
AUC	**A**rea **U**nder **C**urve
SSD	**S**ingle **S**hot Multibox **D**etector
HU	**H**ounsfield **U**nit
POI	**P**oint **O**f **I**nterest
BB	**B**ounding **B**ox
DPN	**D**ual **P**ath **N**etwork
NMS	**N**on **M**aximum **S**uppression

Chapter 1

Introduction

Lung cancer is the leading cause of cancer deaths, estimating 2 million new cases yearly [4]. A patient diagnosed with lung cancer could potentially be treated through procedures such as chemotherapy, immunotherapy and other systemic anti-cancer therapies. Traditionally radiologists and doctors look at computed tomography (CT) scans to recognize and diagnose lung cancer tumors or cancerous nodules. However, such methods are time-consuming, expensive and difficult to scale [18]. Many automatic diagnostic systems have been developed to predict cancerous pulmonary nodules using CT scans. With the development of deep learning, a number of researchers are trying to adopt deep learning models for this purpose. These systems must be highly accurate as they are used for clinical diagnostics and analysis. However, deep learning methods require a large amount of training data to attain desirable results while size of available medical datasets is small [14].

For annotating large amounts of clinical data, we require medical experts. However, it is difficult and expensive to manually annotate a large-scale medical dataset. There have been several studies indicating the feasibility and reliability of using non-experts via crowdsourcing for large-scale image annotation tasks [12]. Due to the success of crowdsourcing, many researchers have applied crowdsourcing for medical image analysis. Although medical images present several challenges like task complexity and lack of domain knowledge required by workers, these works have shown promising results[26] [3].

This project aims to answer the following research questions: (i) Can crowdsourcing using non-experts be used to annotate a large scale lung CT dataset for nodule detection? (ii) Can such crowd engineered pipelines refine the performance of state of the art deep learning models

or nodule detection? (iii) What is the optimal combination of human intervention and machine computation to achieve high accuracy?. Due to the cost effectiveness and success of crowdsourcing methods for annotation tasks, we propose an approach that uses non-experts via crowdsourcing to annotate lung cancer tumors. We will feed these annotations to an automated tumor detection algorithm, then evaluate the extent to which we can improve its accuracy as shown in Figure . We hypothesize that annotations collected from non-experts using crowdsourcing will be able to refine and enhance the existing state of the art deep learning models.

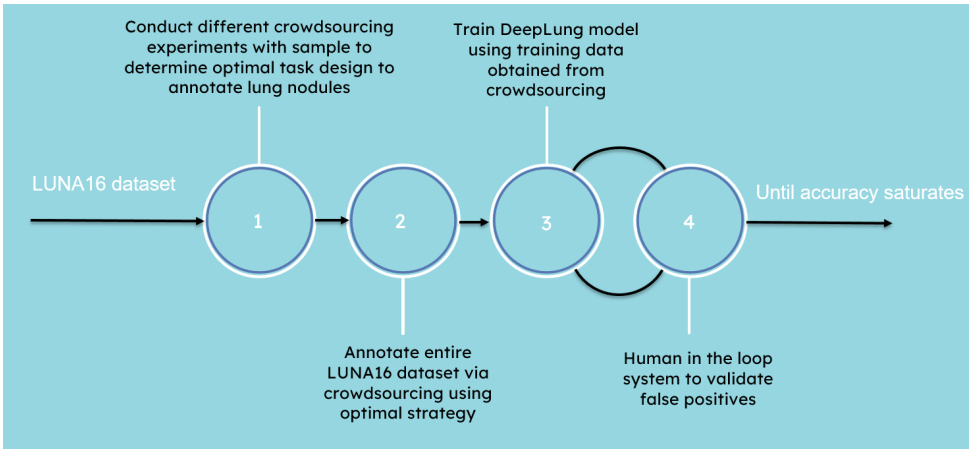

FIGURE 1.1: The proposed architecture for the project. In stage 1, different crowdsourcing experiments are conducted to determine optimal strategy. Stage 2 uses the optimal strategy for annotating LUNA dataset by crowd. Stage 3 trains the DeepLung model using crowdsourced annotation. Stage 4 iteratively refines the DeepLung model using human in the loop system.

For evaluation, we use our crowd-annotated dataset to train the DeepLung model [13], which has already produced high accuracy (90.44% nodule classification accuracy) with an expert-annotated dataset. Since our crowdsourced training dataset may have a high number of false positives, we use a human-in-the-loop system to correct these mislabeled nodules and iteratively carry this process until accuracy reaches the maximum level.

Thus, this project aims to create a system combining human intelligence and machine learning capabilities to enable the creation of low-cost datasets for automatic diagnostics and human-machine integrated systems to reduce the problem of high sensitivity of Computer Aided Diagnostic Systems.Our main contributions are as follows: i.) Design a crowdsourcing methodology for medical images so that the annotations collected from the crowd have high correlation with the ground truth. ii.) Refine existing state of the art deep learning models for tumor detection using crowdsourced annotations and active learning.

Chapter 2

Related Work

We discuss the related work for this project from three perspectives, namely: i.) crowdsourcing medical annotations especially involving with lung CT images, ii.) automatic tumor detection systems using deep learning models and iii.) active learning systems to refine deep learning models for medical diagnosis.

2.1 Crowdsourcing

Crowdsourcing has shown promising results for large-scale annotations for object detection tasks in natural images.[12]. Due to the success of crowdsourcing, many researchers have applied crowdsourcing for medical image analysis. Although medical images present several challenges like task complexity and lack of domain knowledge required by workers, these works have shown promising results [26]. *Ørting et al.* [26] surveys the recent research using crowdsourcing for both classification and localization for medical images. The work outlines application, task design, scale and evaluation metrics implemented in the reviewed research. This survey suggests that research pertaining crowdsourcing for lung images is still in its infancy. *O'Neil et al.* [21] proposed to use non-experts to annotate pathological regions associated with interstitial lung disease. Participants received an hour-long training and were assigned 2D scans in random order and asked to draw a polygonal ROI [ix]. Tasks were repeated in 10 days to check for reproducibility of the assignment. *Boorboor et al.* [1] used crowdsourcing to find cancerous nodules in lung CT scans. The crowdsourcing task was presented as a TS-MIPS [ix] video of 5 consecutive slices which the crowd could pause and annotate when encountered with a tumor.

90% nodule detection sensitivity was obtained with only 47 FPs [ix] from 1021 annotations. *Cheplygina et al.* [5] used crowdsourcing to annotate airways in the lung CT scans. Even though, after post filtering many annotations were deemed unusable, the final aggregated annotations had medium to high correlation with expert annotations. Even though the results from these works are promising, the experiments were carried out on a small scale of 20 scans . We aim to use crowdsourcing to annotate lung nodules in CT scans on a larger dataset and compare different experimental designs to obtain high accuracy. To reduce the number of unusable annotations, we will carry out two rounds of crowd annotation tasks on the same image as indicated in Chapter [3].

2.2 Automated tumor detection systems

2.2.1 Image Processing based systems

We aim to use the annotated data from the crowdsourcing experiments for training a deep learning model to detect lung cancer tumors. Early research in this field used traditional image processing techniques to segment tumors from lung 2D CT scan images. *Makaju S., Prasad et al* [16] used Median and Gaussian filters for image enhancement with watershed segmentation technique for tumor segmentation and feature extraction. *Uzelaltinbulat et al.* [22] created a three-stage segmentation system using Otsu segmentation technique. Various filters were applied for image smoothing, boundary removal etc. However, these techniques can only be used for a few scans and are not scalable. In addition, since tumors vary in size, density and orientation to a great extent, it is very difficult to identify tumors accurately using traditional techniques alone. Another drawback is that CT scans are innately 3D in nature and information is lost by not utilizing this factor and analyzing 2D images independently.

2.2.2 Deep Learning based systems

With new advances in deep learning for natural images, researchers are trying to replicate these architectures for medical images to obtain better accuracies. Recently, several works have been proposed to use Deep Convolutional Neural Networks for nodule detection to automatically learn features, which has proven to be more effective than hand-designed features. Deep learning methods enabled the CAD [ix] system to be implemented as a single stage process without

having to do image smoothing and filtering. *Naji K., Ulas Bagci* [19] implemented a Single shot single stage Lung Nodule detection technique using 3D convolutional network and cross entropy for cost optimization. The model gave a sensitivity of 95.2% and a FROC [ix] score of 0.897 performing better than SSD [ix], which is the state of the art object detection algorithm. *Chon A., Balachandar N.* [6] built a generalized model for lung nodule detection using a modified U-Net structure implemented using 3D CNNs for tumor segmentation and a 3D CNN structure similar to Google Net for classification. *Hossain S. et al* [9] developed a CAD system by using dilated 3D CNNs based on Lung Net architecture. *H. Tang et al.* [8] built a 2 stage network using U-Net inspired by Faster RCNN and hard mining and a 3D DenseNet CNN model for the classification stage. *M. Winkens, Cohen* [23] using 3D group CNNs using data augmented using rotation and reflection in order to create more data for better performance of the CNN network. *Zhu W. et al* [25] created DeepLung using 3D dual path networks with a Faster RCNN inspired UNet and a classification structure to identify tumors. This model was compared with the 3D CNN model using ResNet[11] model and performed better than the model using 1/4th the number of hyperparameters. Due to the effective performance of such 3D Faster R-CNN based UNet for nodule detection and a 3D CNN for nodule classification; we will use a similar design for our deep learning model.

2.3 Active Learning

X. Xie et al. [13] uses an alternative approach inspired by active learning for breast cancer detection. This research proposed to train the deep learning model on overlapping patches of a biopsied slide wherein all patched receive the same label as the entire slide. This leads to mislabeled patches, the problem which they solve using Reverse Active Learning. The model is initially trained on the entire dataset, a human in the loop validates the mislabeled patches predicted by the machine, and this experiment is carried out for multiple iterations until the accuracy saturates. Since our crowdsourced annotations may have a large number of false positives,a similar approach is used with a human in the loop to validate the misclassified nodules and iteratively improve the deep learning model.

Chapter 3

Methodology

3.1 Dataset

In this project, the LUNA16 dataset [15] dataset is used. Though this dataset are already annotated by experts, crowdsourcing is used to annotate the dataset to answer the question of whether 'non-expert' crowd is able to annotate these datasets. The expert annotations provided by the dataset are used as a gold standard to evaluate the crowdsourced annotations.

The LUNA16 [15] dataset's annotations consists of the center of the tumor and its diameter width with respect to the 3D .mhd file. The LUNA16 subset is a subset of the LIDC-IDRI(1) [2] dataset which consists of 1000 CT scans. Scans less than 2.5 mm in the LIDC-IDRI have been excluded. Overall, the dataset comprises of 888 total patient scans with 1186 labelled nodules annotated by experienced radiologists. The LIDC-IDRI database was curated by 4 expert radiologists in which annotations were collected in a two-phase process. Each radiologist marked lesions[1] as not a nodule, nodule with diameter width <3mm and nodules >= 3mm. In the LUNA16 dataset, only nodules greater than 3mm which were accepted by at least 3 out of 4 radiologists were included. A subset of the LUNA16 is used for curating the training dataset for crowdsourcing and the remaining data is used for validation and testing.

[1] lesion: A region in an organ or tissue which has suffered damage through injury or disease, such as a wound, ulcer, abscess, or tumour

3.2 Crowdsourcing

The concept of crowdsourcing was first coined by Jeff Howe, in 2006, as "the act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people of an open call" [10]. Thereupon, crowdsourcing methods have been applied in numerous fields (object localization and classification, entity relations, text annotations etc.) since it is cost-effective, efficient and often times results in expert-like accuracy. The main task for the workers in this project is to locate and classify a structure as a tumor. An important aspect of crowdsourcing is the task design. The relation between type of image data, annotations required and available tools plays an important role in quality of results obtained [26]. In order to retrieve the best possible annotations from the crowds, the images are pre-processed such that the task is less complex for non-experts. The detailed pipeline of crowdsourcing experiments is shown in Figure 3.1.

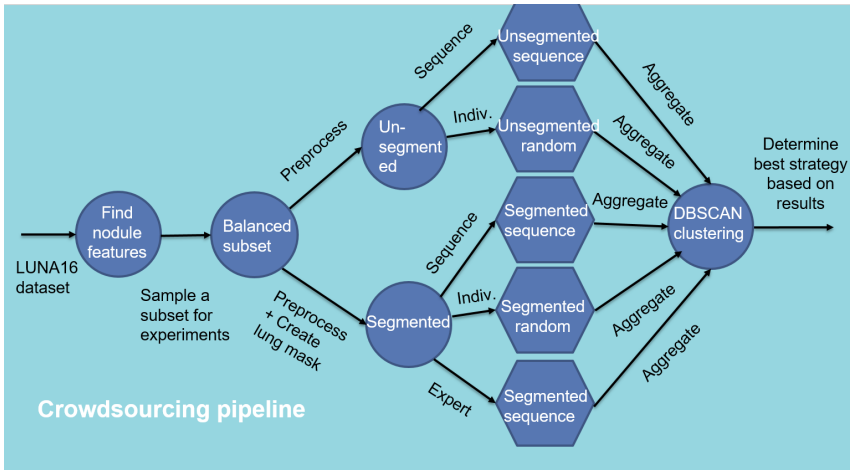

FIGURE 3.1: Detailed pipeline of crowdsourcing experiments as part of stage 1 of the project.

3.2.1 Preprocessing for crowdsourcing experiments

The overall preprocessing process is shown in Figure 3.2. For the crowdsourcing experiment, the data from the LUNA dataset is used. The annotations in the LUNA dataset are given in terms of world coordinates while to locate a tumor, the 3D voxel coordinates must be determined. Therefore, the origin of the scan and spacing between slices are retrieved using SimpleITK library [2]and the tumor coordinates are converted from world to 3D voxel coordinates. For the crowdsourcing experiment, we selected a balanced subset comprising of equal number of positive

[2]`https://github.com/SimpleITK/SimpleITK`

samples (tumor-containing) and negative samples (non-tumor containing) for human intelligence tasks (HITs). However, a 3D CT scan consists of approximately around 200 or more 2D slices. Including all the slices in the experiment creates an imbalanced dataset and increases the scale and cost of this task. Therefore for the positive samples, only five slices adjacent to the tumor center containing slice, are included for crowdsourcing experiment. To balance the sample, an equal number of candidate non-nodules which are mentioned in the LUNA dataset are included in the sample for crowdsourcing experiment.

Since the scans in the dataset are in the axial view, other organs like the thorax, rib cages and the upper abdominal region are also captured in the lung CT scan along with the lung parenchyma. Non-experts can mark any of these regions of a CT scan as they do not possess the technical knowledge of analyzing a chest CT scan. In order to avoid causing such errors, we make use of a lung mask [16][22] using image processing algorithms to remove all the parts other than the lung parenchyma. To generate the lung masks, all the pixels in each CT slice are normalized and are loaded as a 512*512 numpy array. A thresholding technique is used to obtain the lung mask. To obtain the optimal threshold, KMeans clustering (n=2) is used to separate the foreground and the background of the image (Soft tissue have a high contrast difference as compared to the thorax, ribs and diaphragm). Once the thresholded image is obtained, erosion and dilation operations are carried out to eliminate noise and improve segmentation. Since the lungs typically lie in the center of the scan, all the components on the top (<[0, 40]) and bottom (>[475,512]) are removed. A final dilation is performed to fill in and out the lung mask. The segmented image is obtained by multiplying the lung mask obtained above with the original image. This process results in segmented lung CT scans. The experiments conducted for crowdsourcing tasks contain both segmented and unsegmented images.

(A) (B) (C)

FIGURE 3.2: (a) 2D slice of the lung CT scan which is unsegmented .(b) Extracted lung mask after applying thresholding and morphological operations.(c) Segmented lung parenchyma.

To increase visual ease in comprehending between different organs and other biological elements of a CT scan, the raw data is clipped between [-1000, 400] and data is linearly transformed between [0, 1]. This range of values in Hounsfield Unit (HU) [3] ([-1000, 400]) are recommended by radiologists and is known as lung settings [3]. The images are saved in a lossless .png format after applying a gray cmap filter for the crowdsourcing experiment.

3.2.2 Experiments

The workers for this task are chosen from Amazon Mechanical Turk [4] platform, wherein they are redirected to our custom annotation interface as shown in Figure 3.4. In order to train non-experts about the physical appearance of a tumor and the possible locations, a tutorial is presented for the worker which interactively gives feedback in case an incorrect structure is marked as a tumor. To determine an optimal design strategy for CT scan images to be accurately annotated, multiple experiments are carried out. Since multiple experiments are performed, a representative subset is sampled from the LUNA16 dataset.

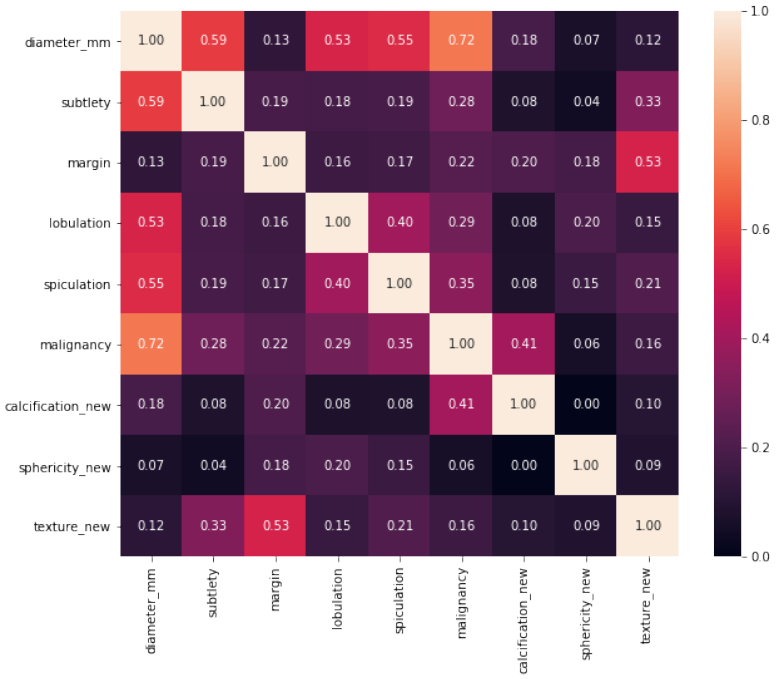

FIGURE 3.3: Correlation heatmap of nodule characteristics in the LUNA16 dataset

[3] The Hounsfield Unit (HU) is a measure of radiointensity for CT scans. It is based on the amount of absorption of radiation within a tissue and is used in CT restoration to generate a grayscale image.

[4] `https://www.mturk.com/`

3.2.2.1 Sample selection

The LUNA16 dataset consists of 888 lung CT scans with 1186 nodules which are classified as tumors. These tumor nodules vary with respect to diameter from 3mm to 30mm. To select the sample subset for the experiments, the distribution of physical nodule characteristics is considered. The LIDC-IDRI dataset, which is the superset of the LUNA16 dataset, consists of radiologist annotated metadata regarding the nodule characteristics. Characteristics like subtlety, internal structure, calcification, sphericity, margin, lobulation, spiculation, texture and malignancy are rated on a scale of 1 to 5 depending on the physical appearance by radiologists [2]. These characteristics are included in the XML annotations of the LIDC dataset as additional metadata. All these characteristics are retrieved from XML annotation files and mapped to their corresponding nodules in the LUNA16 dataset. From the retrieved nodule characteristics, the correlation is calculated to determine the most significant features for the tumor distribution. The correlation heatmap produced using Cramer's V correlation for nominal variables is shown in Figure 3.3. From this heatmap, diameter width of the tumor is observed to be the most influential parameter for tumor distribution. Therefore, a sample of 20 nodules is selected using stratified random sampling [5] based on the nodule's diameter distribution as the positive sample from a population. Similarly, a subset of 20 non-nodules is selected as the negative sample. Therefore, the final test set consists of 40 nodules, each nodule consisting of 5 consecutive tumor slices around its center amounting to 200 test images for the crowd.

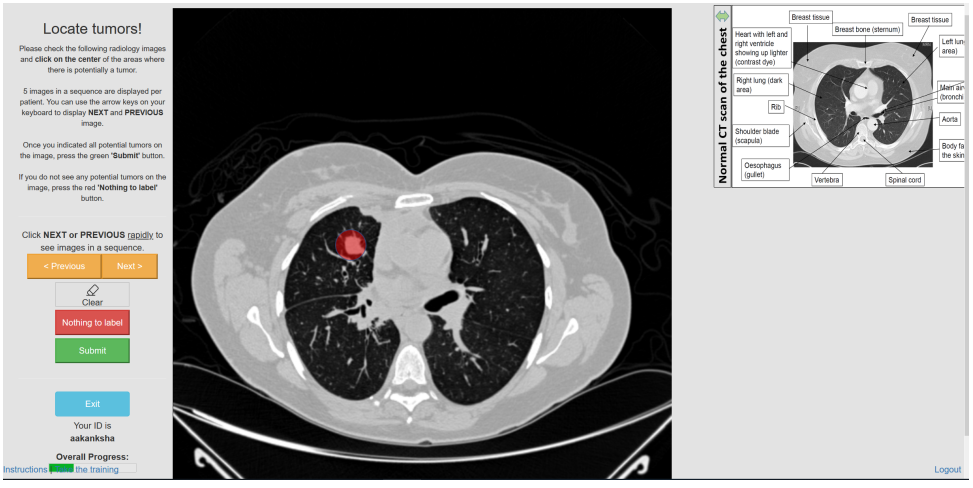

FIGURE 3.4: Task presentation in the custom application for annotating lung tumors by the crowd

[5]Stratified random sampling: Random sampling performed on population subgroups or strata divided based on a common characteristic of the group

Type of Annotation	Point of Interest (POI)	Bounding Box	Contouring
Task Difficulty	Lowest	Medium	Highest
Time/Cost	Lowest	Medium	Highest
Prone to Error	Highest	Medium	Lowest

TABLE 3.1: Comparison of types of annotation

3.2.2.2 Experimental design for crowdsourcing

The key variables for the presentation of tasks are i.) Type of image ii.) Batch/Static Presentation and iii.) Type of annotations. 3D visualization and annotation requires medical expertise which crowds do not possess. Therefore, scans are presented as 2D slices where the crowd can identify the 2D projection of the tumor and annotate them. With respect to type of image, an image can be segmented (from the extracted lung mask) or unsegmented as shown in Figure 3.2, Section 3.2.1. Therefore, 2D slices are presented in two formats: i.) Presenting a batch of 5 consecutive 2D CT slice together as a microtask (may or may not contain a tumor) which we call sequence microtasks ii.) Randomly presenting 2D slices as static standalone microtasks (adjacent images are not consecutive).

Based on the key features identified, we have designed four different experiments. The four experiments include (i) Segmented sequence experiment, (ii) Unsegmented sequence experiment, (iii) Segmented random image experiment (iv) Unsegmented random image experiment. Few examples of the scans used in the experiments are shown in Appendix A.1. Each of these experiments is a "Find" experiment which uses the point of interest annotation to mark the tumor centers on all the possible slices where tumor is visible on the unlabelled image data. Once a point is marked, the blob sequence in consecutive slices is used to find the tumor based on the point's position. The experiments are annotated using point of interest annotation as it is cheaper with respect to task difficulty and time/cost expense as shown in Table 3.1 . Even though it is prone to error more than a bounding box or a contouring approach, majority voting aggregation method takes care of this problem as we assume that the majority of the crowd will annotate the correct region.

For the experiments involving microtasks as image sequences, five consecutive slices were selected from a patient, while for the randomly presented microtasks, one image was shown from any patient in the dataset.For each experiment, 20 annotators were selected from Mturk. The

Number of subtasks	40
Number of subtasks containing a tumor	20
Number of annotators for each experiment	20
Type of annotation	Point of Interest (clicking on the center of tumor)
Dimension of Image in task	2D
Reward for completed task	$3 (partial reward depending on proportion of completed task)
Number of annotations per subtask by a single worker	1
Segmentation	Only Lung Parenchyma present, other structures from the image are discarded

Experiment No.	**Type of Image in subtask**	**Presentation of images in subtask**	**Number of images in a subtask**
Experiment1	Unsegmented	Sequence of consecutive CT slices from same patient	5
Experiment2	Segmented	Sequence of consecutive CT slices from same patient	5
Experiment3	Unsegmented	Random CT slices from any patient	1
Experiment4	Segmented	Random CT slices from any patient	1

FIGURE 3.5: Details for crowdsourcing experiments conducted to determine optimal strategy to retrieve annotations from crowd

experiments were conducted on a custom annotation interface designed [6] for this project to annotate lung nodules as shown in Figure 3.4 [7].

Each task consists of a balanced set of 20 nodules containing microtasks (positive examples) and 20 non-nodule containing (negative examples) microtasks. For clarity, the task mentions that, if a tumor is present, then there is only one tumor to annotate per microtask. Incase the worker encounters a microtask with no tumor, then the worker must click on 'Nothing to Submit'. The worker could exit in the middle of the task. The compensation for participating in this task was $ 0.5. To stimulate crowd motivation, a bonus of upto $ 3 was awarded depending on the whether task was completed and quality of annotations. The details about the experiments is provided in Figure 3.5.

To assess the difference between using 'experts' and 'non-experts' for annotating nodules in CT scan slices, we repeated the experiment with a 'expert' group. This group comprises of scientists working in the field of biomedical engineering and cancer diagnostics, who volunteered to do the same exercise as the crowd. This experiment will help us to understand if there is a considerable difference in performing experiments with 'non-experts' as compared to 'experts'.

[6] `http://kitlekaynak.net/mia/medical_poi/main.php`
[7] `https://drive.google.com/file/d/1igbyOIBOPtN7jzGYfJREMyukawCez9yk/view?usp=sharing`

Chapter 4

Results and Discussion

4.1 Crowdsourcing

The crowdsourcing experiments included subtasks implementing four different strategies; (i) Segmented sequence presentation, (ii) Unsegmented sequence presentation, (iii) Segmented random presentation and (iv) Unsegmented random presentation [refer Section 3.2.2.2]. These images were annotated by the crowd using point of interest annotation. The segmented CT scan slices were generated by applying the preprocessing steps as mentioned in the methodology 3.2.1.


```
    "id": "5dcacec4dc861",
    "annotator": "banu",
    "url": "input\/CROWD_TEST\\Negative_samples\\seg_op\\subset1\\1.3.6.1.4.1.14519.5.2.1.6279.6001.20653988515
477500292903153429 1\\124\\.png",
    "timestamp": "2019-11-12 16:24:52",
    "POIs": "[{\"x\":101,\"y\":444}]",
    "Polygons": "[]",
    "FreeDots": "[]",
    "Boxes": "[]",
    "DisplayTime": "1573572287386",
    "SubmitTime": "1573572292899"
```


FIGURE 4.1: Annotation parameters collected for an image on our custom annotation interface in json format

The annotations collected from our custom interface records a number of parameters regarding the annotation in JSON format as shown in Figure 4.1. Each annotator was assigned a unique identifier which could be submitted on Mturk to receive the payment for the task. The url specified the file name of the image for which the worker created the annotation. Each annotation recorded the time taken for the worker to annotate the subtask. Annotations were collected as x and y coordinate of the center of the potential tumor. To ensure that the annotator performed

the task carefully, the average time (t_{worker}) taken per subtask was calculated. If t_{worker} lies in the range of [7,100] seconds, the annotation is considered valid. Any value out of this range is discarded[1]. Further, to ensure quality of annotations, the number of submissions(annotated tumor/ Nothing to mark) must be greater than 40 (out of 46) and the number of annotations indicating a tumor is present must be greater than 10.

To test the quality of annotations collected in the experiments, the ground truth annotations from the expert radiologists annotations of the LUNA16 [15] dataset were retrieved for the sample. Nodules were identified based on their patientid, x, y, z coordinates of the tumor's center, diameter width and CT slice number in which the center of the tumor lies. Since the coordinates in the LUNA16 dataset annotations are for a 512*512 CT slice while our custom interface collects annotations for a 800*800 image, the crowdsourced annotations were linearly scaled to 512*512.

4.1.1 Worker wise performance for crowdsourcing tasks

For all the workers who had quality annotations, the true positives (TP), true negatives (TN), false positives (FP) and the false negatives (FN) were calculated. A true positive (TP) annotation is defined as a point which lies within 50 pixels of ground truth centroid point of a tumor (50 pixels were considered since tumor sizes range from [3,32] mm). A false positive (FP) is defined as a point marked where a tumor is not present or a point lying at a distance of more than 50 pixels. If an image from the negative sample of the subset is submitted as a 'Nothing to submit', then it is considered as a true negative (TN). If a tumor containing slice is marked as 'Nothing to submit', then it is considered as a false negative (FN). Once all the annotations are classified as either TP, TN, FP or FN; the recall, precision and the true positive rate are calculated per worker to evaluate worker-wise performance.

4.1.1.1 Experiment 1: Unsegmented sequence subtasks

The assignment of annotations for the unsegmented sequence experiment is shown in Table 4.2. 11 out of 20 annotators passed the quality check established for the project. Overall, the number of false positives (FPs) for the experiment is high (mean FP=24/41) while the number of true positives is much lower (mean TP=9/41). Based on the results, it was observed that even if

[1]This is the average time bracket which is estimated based on the worker performance in a particular experiment, bracket may be shifted according to the each experiment's obtained estimates

outcome annotator	FN	FP	TN	TP	recall	precision	TP_Rate
A10Q4U3BRHXXPP	0.0	27.0	0.0	14.0	1.000000	0.341463	0.341463
A153J31AVDX32V	0.0	35.0	0.0	6.0	1.000000	0.146341	0.146341
A1GKD3NG1NNHRP	11.0	17.0	8.0	5.0	0.312500	0.227273	0.121951
A23PQYQ6A2I076	4.0	33.0	2.0	2.0	0.333333	0.057143	0.048780
A2CYXHEA1EX07O	0.0	36.0	0.0	5.0	1.000000	0.121951	0.121951
A2M1CVZZJAN4T4	0.0	21.0	3.0	17.0	1.000000	0.447368	0.414634
A31XFBQITA3FAP	0.0	32.0	3.0	6.0	1.000000	0.157895	0.146341
A3F8UT6178B2A4	3.0	10.0	13.0	15.0	0.833333	0.600000	0.365854
A3UF6XXFFRR237	1.0	28.0	0.0	12.0	0.923077	0.300000	0.292683
ACI8PUCF5OPDC	2.0	27.0	4.0	8.0	0.800000	0.228571	0.195122
AWENQ6RS7ABZ6	7.0	5.0	16.0	13.0	0.650000	0.722222	0.317073

FIGURE 4.2: Worker performance for unsegmented sequence experiment (considering only quality annotations)

the recall for certain workers is 1.00, the precision is fairly low. This is due to the fact that these workers marked all the images as tumor-containing while the subset is a balanced one (20 positive and 20 negative samples). This suggests that workers tend to find some position to mark on every image without the considering that the image might not contain a tumor. To understand the worker-wise performance better, we calculate the metrics by considering only the positive samples. In this case, the true negatives (TN) will always zero. The results are shown in Table 4.3.

Removing the negative samples from the subset, approximately doubles the precision and the true positive rate while recall remains the same. This demonstrates the worker-wise performance better. However, due to the problem of high number of false positives (FPs) compared to the true positives (TP), precision is moderate (mean precision=0.48). This can theoretically be remedied by including higher number of annotators per image.

For the rest of the experiments, we will compare only the worker performance on the positive samples of the subset.

4.1.1.2 Experiment 2: Segmented sequence subtasks

The assignment of annotations for the segmented sequence experiment is shown in Table 4.4. 10 out of 20 annotators passed the quality check established for the project. For the positive samples (20 subtasks), the number of false positives (FPs) for the experiment is high (mean FP=13/20) while the number of true positives is much lower (mean TP=6/20). This experiment

outcome annotator	FN	FP	TP	recall	precision	TP_Rate
A10Q4U3BRHXXPP	0.0	8.0	13.0	1.000000	0.619048	0.619048
A153J31AVDX32V	0.0	15.0	6.0	1.000000	0.285714	0.285714
A1GKD3NG1NNHRP	11.0	6.0	4.0	0.266667	0.400000	0.190476
A23PQYQ6A2I076	4.0	15.0	2.0	0.333333	0.117647	0.095238
A2CYXHEA1EX07O	0.0	16.0	5.0	1.000000	0.238095	0.238095
A2M1CVZZJAN4T4	0.0	5.0	16.0	1.000000	0.761905	0.761905
A31XFBQITA3FAP	0.0	15.0	6.0	1.000000	0.285714	0.285714
A3F8UT6178B2A4	3.0	4.0	14.0	0.823529	0.777778	0.666667
A3UF6XXFFRR237	1.0	9.0	11.0	0.916667	0.550000	0.523810
ACI8PUCF5OPDC	2.0	12.0	7.0	0.777778	0.368421	0.333333
AWENQ6RS7ABZ6	7.0	2.0	12.0	0.631579	0.857143	0.571429

FIGURE 4.3: Worker performance for unsegmented sequence experiment with only positive samples (considering only quality annotations)

	Worker ID	FN	FP	TP	Recall	Precision	Accuracy
0	A1C0AWJ1JULKO2	0	10	10	1.000000	0.500000	0.50
1	A24JKHC4HTY6CD	0	11	9	1.000000	0.450000	0.45
2	A3CASN6JG7104	2	14	4	0.666667	0.222222	0.20
3	A3SM8VVB534E7Q	0	17	3	1.000000	0.150000	0.15
4	A3UF6XXFFRR237	0	6	14	1.000000	0.700000	0.70
5	AO2WNSGOXAX52	0	15	5	1.000000	0.250000	0.25
6	AUMTP6BXBDBXL	7	4	9	0.562500	0.692308	0.45
7	GAJA	0	20	0	0.000000	0.000000	0.00
8	TERRY	1	17	2	0.666667	0.105263	0.10
9	jaya	1	16	3	0.750000	0.157895	0.15

FIGURE 4.4: Worker performance for unsegmented sequence experiment with only positive samples (considering only quality annotations)

also shows the crowd's bias to mark on every structure closely resembling a tumor, resulting in a high recall and lower precision.

4.1.1.3 Experiment 3: Segmented random subtasks

The assignment of annotations for the segmented random experiment is shown in Table 4.5. 10 out of 20 annotators passed the quality check established for the project. For the positive samples (20 subtasks), the number of false positives (FPs) for the experiment is high (mean FP=14.5/20) while the number of true positives is much lower (mean TP=5/20). This experiment also shows

	Worker ID	FN	FP	TP	Recall	Precision	Accuracy
0	Ashlei	0	16	4	1.000000	0.200000	0.20
1	Brian	0	17	3	1.000000	0.150000	0.15
2	Jhon	1	17	2	0.666667	0.105263	0.10
3	Krish	0	14	6	1.000000	0.300000	0.30
4	Nathan	0	12	8	1.000000	0.400000	0.40
5	THOIR THOMS	1	19	0	0.000000	0.000000	0.00
6	jenny	0	14	6	1.000000	0.300000	0.30
7	pri	4	7	9	0.692308	0.562500	0.45
8	rani	0	12	8	1.000000	0.400000	0.40
9	shannon	0	17	3	1.000000	0.150000	0.15

FIGURE 4.5: Worker performance for segmented random experiment with only positive samples (considering only quality annotations)

even greater bias to mark on every structure closely resembling a tumor, resulting in a high recall and very low precision.

4.1.1.4 Experiment 4: Unsegmented random subtasks

The assignment of annotations for the segmented random experiment is shown in Table 4.6. 6 out of 20 annotators passed the quality check established for the project. For the positive samples (20 subtasks), the number of false positives (FPs) for the experiment is high (mean FP=17/20) while the number of true positives is much lower (mean TP=3/20). This experiment fetched very poor results, which shows that presenting microtasks in such a manner is difficult for the crowd to perceive.

UNSEGMENTED - POI						
	FN	FP	TP	Recall	Precision	TP Rate
DEEP	0	19	2	1	0.095238	0.095238
dano	2	18	1	0.333333	0.052632	0.047619
donna	0	16	5	1	0.238095	0.238095
lakiska port	2	18	1	0.333333	0.052632	0.047619
nk	0	18	3	1	0.142857	0.142857
vigil	0	16	5	1	0.238095	0.238095

FIGURE 4.6: Worker performance for unsegmented random experiment with only positive samples (considering only quality annotations)

4.1.1.5 Experiment with experts

To assess the performance of crowd workers, which mainly comprise of 'non-experts' and 'experts', we performed the segmented sequence experiment with a group of 10 scientists with experience in annotating nodules and cancer diagnostics. We selected another group of 10 highly educated scientists but with no domain knowledge. We assume that this group will be more attentive and motivated to do this task than a worker on Mturk. The worker-wise performance with these 'experts' is shown in Table 4.7 . 16 workers passed the quality assessment standard. The average TP for this task was 13/20 and the average FP was 4/20. This shows that 'expert' workers could perform the same task much better than 'non-expert' crowd. This is expected as these experts have the necessary medical domain knowledge required to perform this task which the crowds do not possess.

	Worker ID	FN	FP	TP	Recall	Precision	Accuracy
0	DLAB_Abdalla	3	1	16	0.842105	0.941176	0.80
1	DLab_Esma	0	2	18	1.000000	0.900000	0.90
2	DLab_Lisa	2	2	16	0.888889	0.888889	0.80
3	DLab_Manon	4	7	9	0.692308	0.562500	0.45
4	DLab_Simon	2	4	14	0.875000	0.777778	0.70
5	DLab_Turkey	2	4	14	0.875000	0.777778	0.70
6	Dlab_Guangyao	0	2	18	1.000000	0.900000	0.90
7	Dlab_Sebastian	0	5	15	1.000000	0.750000	0.75
8	IDS_Amrapali	8	4	8	0.500000	0.666667	0.40
9	IDS_Karen	12	2	6	0.333333	0.750000	0.30
10	IDS_Kristi	7	10	3	0.300000	0.230769	0.15
11	IDS_Seun	2	4	14	0.875000	0.777778	0.70
12	Sergey	0	6	14	1.000000	0.700000	0.70
13	Yvonka	4	7	9	0.692308	0.562500	0.45
14	horizon	8	1	11	0.578947	0.916667	0.55
15	rc	10	6	4	0.285714	0.400000	0.20

FIGURE 4.7: Worker performance with experts for segmented sequence experiment with only positive samples (considering only quality annotations)

From the worker-wise performance results for all the experiments, we can see that the sequence tasks perform better than the random static images (Lower FPs, Higher TPs). This shows that presenting subtasks in a batch format makes it easier for the crowd to correctly locate and annotate a nodule. However, in all the experiments, the number of false positives are still higher

than desired as compared to the true positives. Also, the number of workers passing the quality check standard is nearly half the total number of annotators per image. This may be rectified by increasing the number of annotators per image.

4.1.2 Aggregation of crowdsourced annotations per image

Since a number of annotations were collected per image in each crowdsourcing experiment, there must be a aggregation method to locate a point where the crowdsourced annotations agree. For the positive samples, it was observed that the most of the annotations tend to be around the potential tumor location as shown in Figure 4.8. Taking advantage of this factor, DBSCAN [17] (Density-based spatial clustering of applications with noise) algorithm was applied on the set of annotations for each image. Since the number of points are less, atleast two points must lie in a cluster. From the labels obtained from DBSCAN algorithm, centroid of points belonging to the same cluster were calculated. Points labelled as -1 do not belong to any class and are hence considered as outliers.

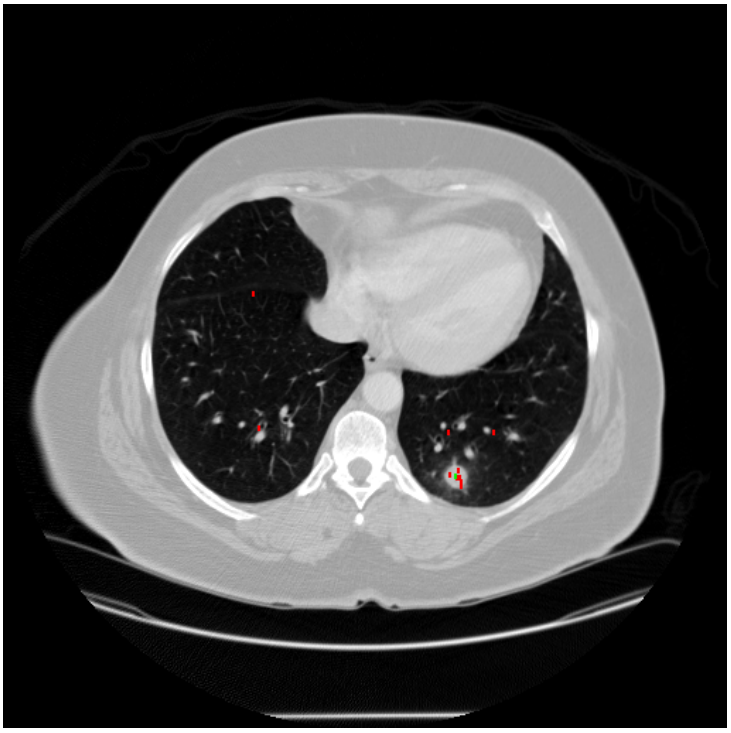

FIGURE 4.8: The annotations marked by the crowd indicating potential tumor locations. Crowd annotations are marked as red points while actual tumor center is marked as a green point.

4.1.2.1 Comparison between aggregated point and ground truth

If no cluster was found for an image, it was assumed that the no tumor is present in the CT slice. For the negative samples, it was observed that the crowd were in disagreement most of the time (Points were labelled as -1). For the negative samples where points were assigned to a

cluster, a structure in the CT scan which closely resembles a tumor was often the cluster centroid as shown in Figure 4.9. The non-expert factor of the crowd is responsible for such clusters as crowds are not familiar with the different anatomical structures in the lung parenchyma. For the positive images, the crowd agreed with the experts most of the time. Although cluster centroids were found in locations where nodules are present as shown in Figure 4.10, many other cluster centroids were also found. This is due to the high number of false positives found in the crowdsourced annotations. There were very few instances in each experiment where aggregation resulted in a positive sample image being marked as one with no tumor. The examples of the types of clusters in the unsegmented sequence experiment is shown in Appendix A.2 .

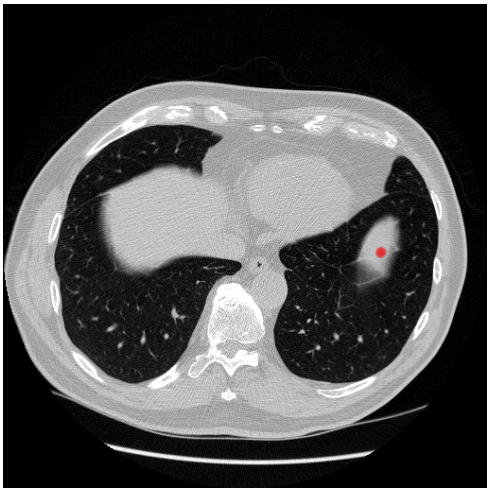

FIGURE 4.9: The red point on the image shows the cluster center generated by agreement of crowdsourced annotations on a negative sample image

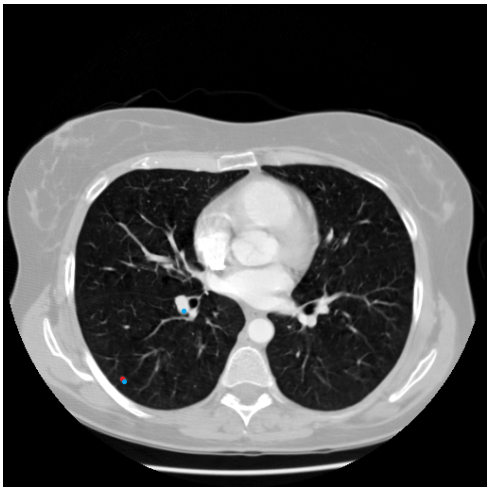

FIGURE 4.10: The blue points on the image shows the cluster center generated by agreement of crowdsourced annotations on a positive sample image. The red point represents the ground truth location of the nodule.

From the annotations recorded per image in each experiments, we calculated the cluster centroids produced and measured how many clusters coincided with the ground truth expert annotations. If the cluster centroid is within 50 pixels of a expert annotation, it is said that the crowd agrees with expert annotation. The detailed results for the experiments with the aggregation is shown in Appendix A.3.

For determining which crowd agreed annotations were best to use, we performed a similar analysis as we did for the worker-wise performance. For each experiment, TP, FP, TN and FN were calculated. Here TP is defined as the number of cluster centroids which coincide with a true nodule in the positive sample. The FP is defined as the cluster centers which do not coincide with any nodule locations. The TN is defined as the number of images in negative sample where no clusters were formed. The FN is defined as the number of images in the positive sample where no clusters were formed. For each of these experiments, the precision, recall and the F1 score was calculated. The results are shown in Figure 4.11.

Experiment	No of cluster centers formed	TP	FP	TN	FN	Recall	Precision	F1 score
Unsegmented Sequence (crowd)	57	21	36	3	2	0.91304	0.3684	0.525
Segmented Sequence (crowd)	29	13	17	9	7	0.65	0.4333	0.52
Segmented random (crowd)	40	8	32	7	1	0.8889	0.2	0.3265
Segmented sequence (Non-crowd IDS)	29	17	12	10	4	0.8095	0.5862	0.68
Segmented sequence (Expert-DLAB)	36	22	14	9	0	1	0.61111	0.7586

FIGURE 4.11: Aggregated annotation results for all crowdsourcing experiments

From the results, we can conclude that the sequence experiments perform better than the random images. After aggregation, the F1 score of the crowd sequence experiments is not significantly different than the expert annotations. This is because the approximate cluster centroids, which we calculate, eliminates the error of an individual point annotation. More annotators per image will further enable the cluster centroids to coincide with the ground truth. Based on all the results, 'Unsegmented sequence experiment' using the crowd is the best strategy to carry the large scale annotation task on the entire LUNA16 dataset (888 scans).

Chapter 5

Conclusion

In this project, we attempted to determine the best design parameters to conduct crowdsourced annotation tasks which can be used in a deep learning model as training datasets. Four different crowdsourcing experiments were performed as part of this project. The experiments involving sequence of consecutive images performed better than static individually presented images. This shows that the crowd benefits from information present in consecutive slices which enables them to identify tumors more accurately than from a single CT slice. Unsegmented sequence experiment emerged as the best performer among all the experiments conducted. We can use this strategy to carry out a large scale annotation tasks involving the LUNA16 dataset. From the results which we obtained, we understand that this is a complex task for the crowd to perform as compared to experts. But when clustering is performed on multiple annotations received by the crowd, the crowd agreement has high correlation with the ground truth.

5.1 Limitations

The biggest limitation of this project is the crowd bias. We used a balanced dataset for the experiments, but the crowd was not notified about the nature of the subset. Since all the images in the training session on the interface had tumors, the crowd assumed that all the images in the test task also will have tumors. This resulted in high number of false positives, where the crowd annotated tumors in CT slices with no tumors. This may be potentially rectified by using a balanced set of examples for the training session as well so that the crowd is aware that non-tumor containing microtasks are also present. Another factor which affected the results was

that training was not mandatory. This resulted in extremely poor performance by some worker, which we assume is affected by not completing the training session. For every experiment, almost half the workers' annotations did not pass the quality check standard set for the experiment. This resulted in lesser annotations than expected. The solution for this problem could be to increase the number of annotators per experiment. Another solution to this problem can be to use the training images as a qualification test. Only high performing workers on the training images will be allowed to annotate on the test task. This will lead to more high quality annotation and will give more accurate cluster centers which coincide with the ground truth.

5.2 Future work

5.2.1 Deep Learning based automatic detection system

Currently, lung CT analysis systems consists of mainly two stages: i.) nodule detection and ii.) nodule classification. Recently, deep Convolutional Neural Networks (ConvNets) such as Faster-RCNN and fully ConvNets [6][8][25][9] are employed to generate potential bounding boxes. Some of these works also use an encoder-decoder UNet model [20] which was developed for precise biomedical segmentation. Because of the effectiveness of Faster-RCNN, a UNet like encoder-decoder model for Faster-RCNN with 3D Convolutional neural networks will be developed for nodule detection to train our crowd annotated CT scans. We will compare the results obtained using the 10-fold cross validation as prescribed in the LUNA dataset using crowd-annotated and expert-annotated datasets to validate our claim.

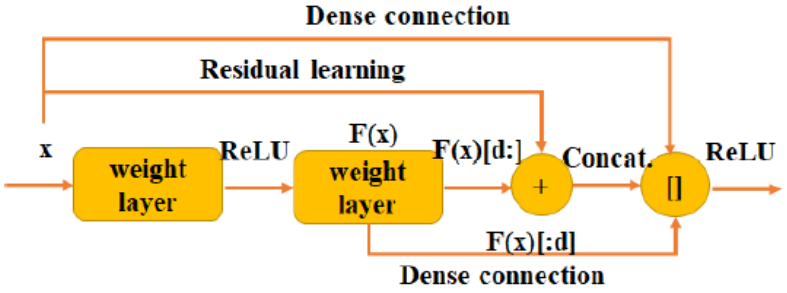

FIGURE 5.1: Dual path connection formed by combining residual and dense connections

5.2.1.1 Model Selection

Due to the success of using dual path networks in DeepLung model [25], we employ a similar architecture by using 3D dual path networks as the base structure for nodule detection and nodule classification. Dual path network (DPN) is a hybrid architecture combining Dense Connected Convolutional Networks (DenseNets) and Residual Networks (ResNets) [24] as shown in Figure 5.1. The shortcut connections in ResNets are an effective method to eliminate the vanishing gradient problem [11]. ResNets allow feature reuse while DenseNets exploit new features [24]. DenseNets use fewer parameters than ResNets since the need to learn redundant feature maps again is eliminated [7]. A dual path network splits part of its features for ResNets and the remaining part for DenseNets. The feature maps $\mathbf{F}(\mathrm{x})[d:]$ are used to represent ResNets while $\mathbf{F}(\mathrm{x})[:d]$ is used to represent the DenseNet feature maps where d is the hyper-parameter to determine how many features to exploit. The dual path connection can be identified as :

$$y = G(x[:d], F(x)[:d], F(x)[d:] + x[d:])$$

where y is the feature map generated for dual path connection, G is the activation function (typically reLU), F is the convolutional layer function and x is the input to the DPN [ix]. We employ this DPN structure for our model due to the feature reusability and effectiveness.

5.2.2 Active Learning

As we have seen in the results of the crowdsourcing experiments, the annotations produced by the crowd have high number of false positives. If we use these annotations to train the deep learning model for nodule detection, the number of false positives for the test set will be high and it will give poor results. To refine the model and improve the quality of our crowdsourced dataset, we will use a human in the loop system to accurately validate and reannotate the misclassified patches by the model. This will be done by setting a threshold to the probability of the patch having a nodule, which is one of the outputs of the nodule detection model. As per the paper by Xie [13], we expect that if this process is carried out iteratively, the quality of the dataset will improve and so will the accuracy.

We will attempt to modify our interface to eliminate the limitations which we faced in this project and construct the deep learning model to use the crowdsourced annotations.This will

enable us to construct a end-to-end crowd engineered deep learning model which will help to reduce the problem of high false positives of computer aided diagnostic systems (CAD).

Appendix A

Appendix

A.1 Experiments for crowdsourcing

To determine the optimal design parameters to conduct a large scale annotation task on lung CT scans, four different experiments were conducted using crowdsourcing. The way the subtasks were presented to the crowd in these experiments is shown below.

A.1.1 Unsegmented sequence experiment

FIGURE A.1: Experiment with unsegmented images in a sequence of 5 slices per subtasks.The slices are retrieved from the same patient and are consecutive

A.1.2 Segmented Sequence experiment

A.1.3 Unsegmented random experiment

A.1.4 Segmented random experiment

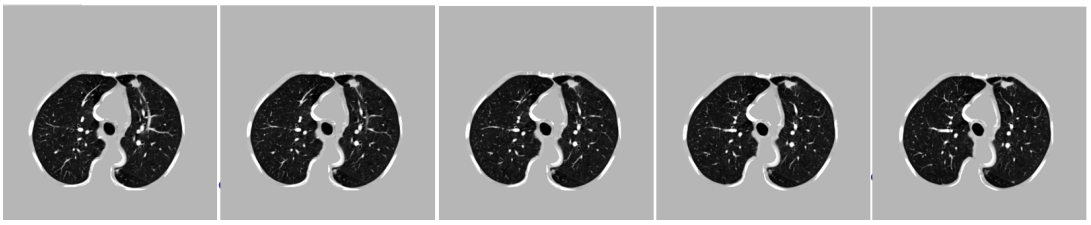

FIGURE A.2: Experiment with segmented images in a sequence of 5 slices per subtasks.The slices are retrieved from the same patient and are consecutive

FIGURE A.3: Experiment with 1 unsegmented image per subtask

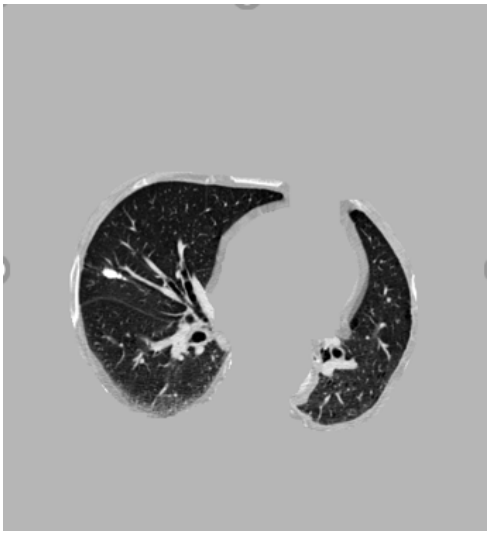

FIGURE A.4: Experiment with 1 segmented image per subtask

A.2 Examples of aggregation of annotations

The red point represents the crowd annotation and the green point represents the ground truth nodule annotation. These images are examples taken from aggregation of unsegmented sequence experiment.

A.2.1 Negative sample image with no clusters (TN)

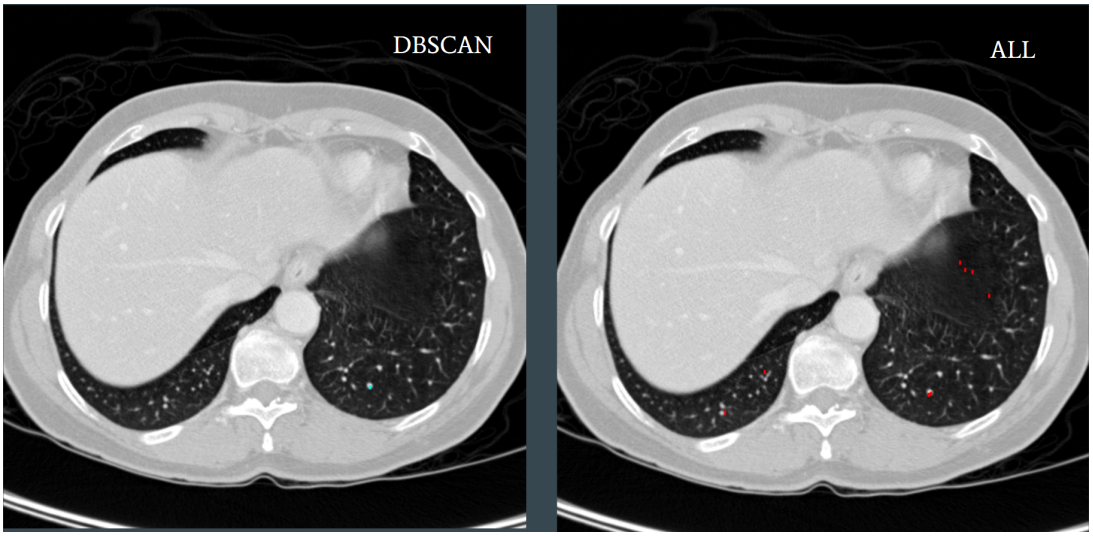

FIGURE A.5: Negative sample image with no clusters (TN)

A.2.2 Negative sample image with clusters (FP)

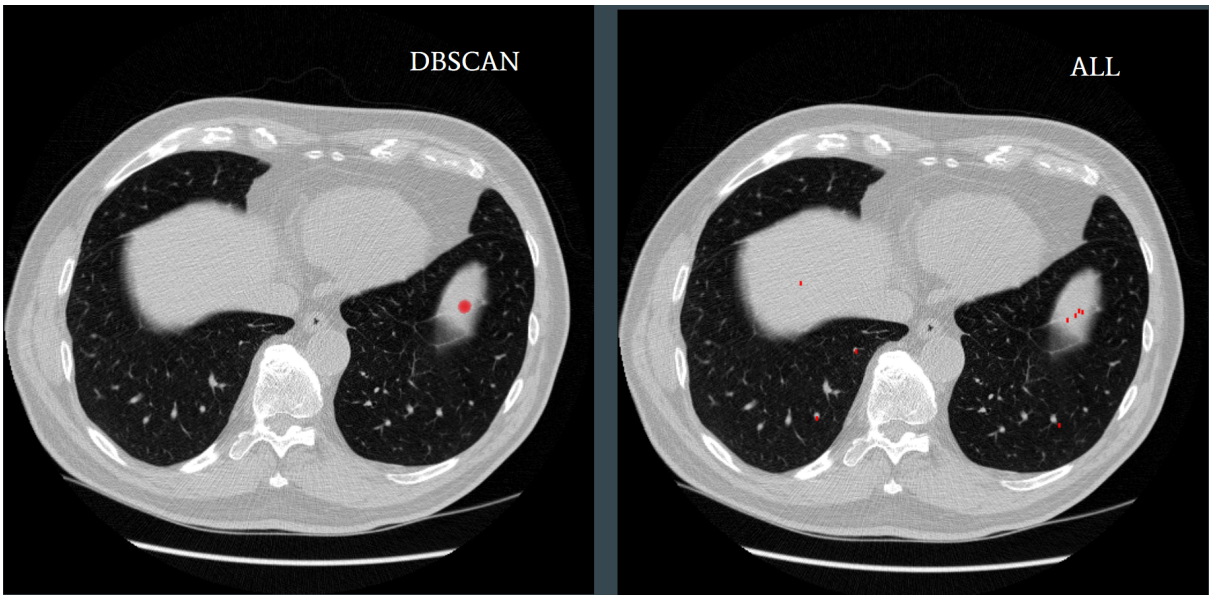

FIGURE A.6: Negative sample image with clusters (FP)

A.2.3 Positive sample image with cluster exactly matching true nodule (TP)

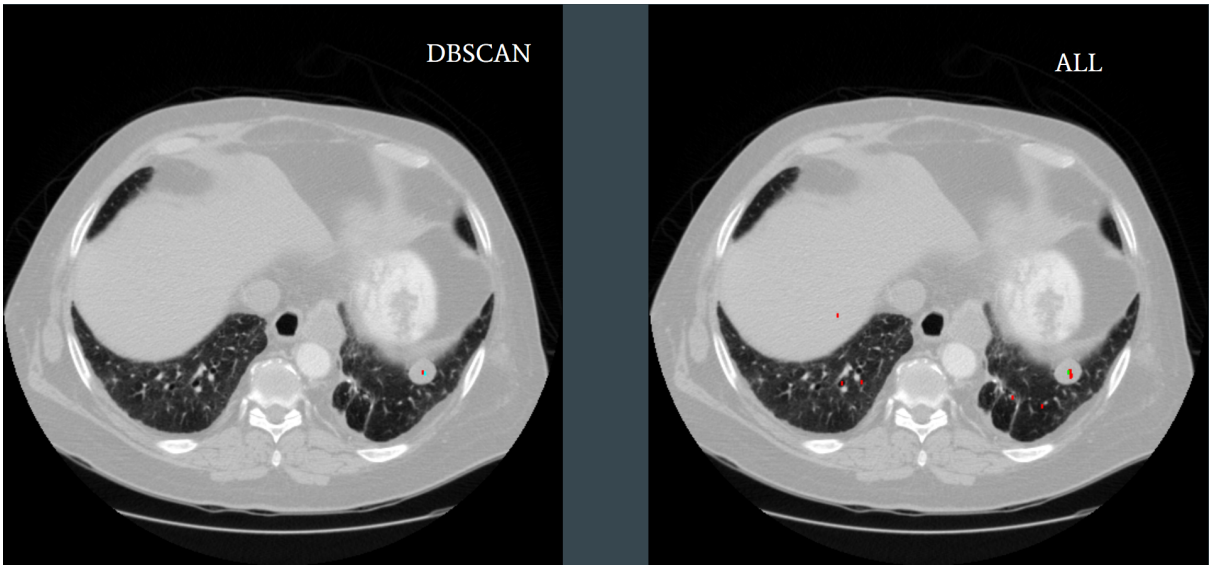

FIGURE A.7: Positive sample image with cluster exactly matching true nodule (TP)

A.2.4 Positive sample images with both TP and FP

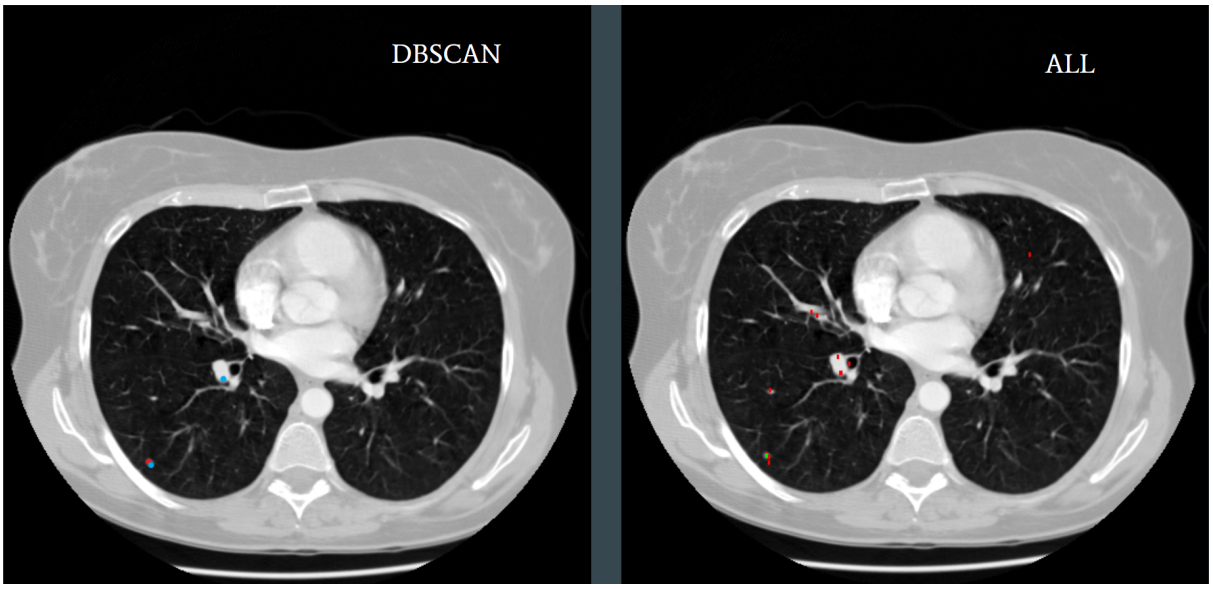

FIGURE A.8: Positive images with both TP and FP

A.2.5 Positive sample image with no cluster (FN)

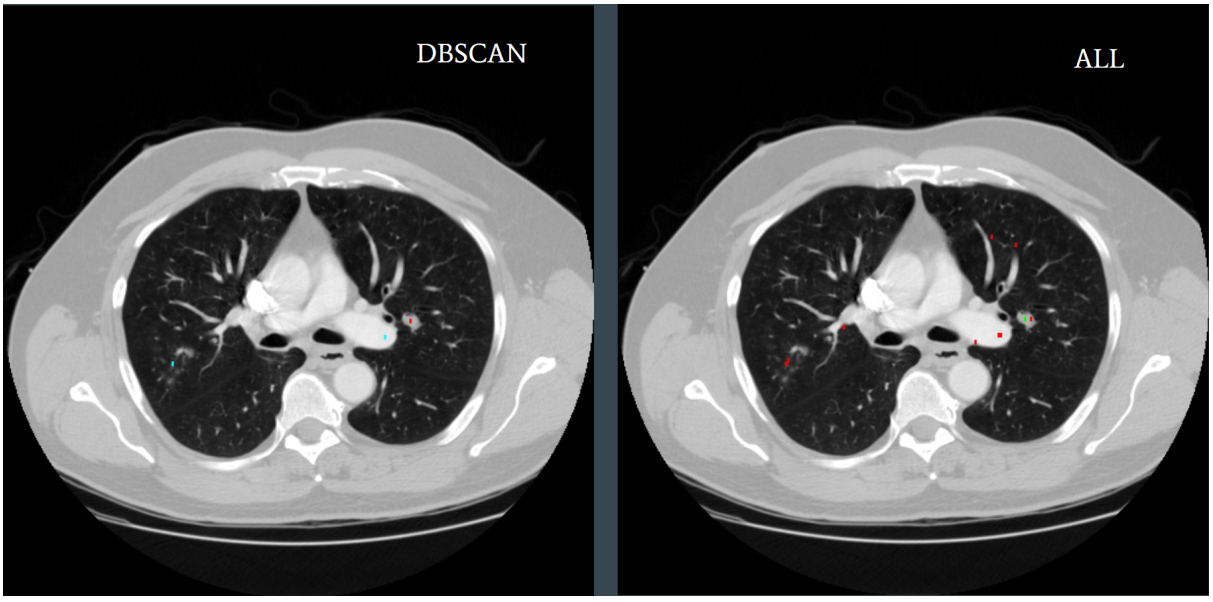

FIGURE A.9: Positive sample image with no cluster (FN)

A.3 Aggregation results for crowdsourced experiments using DBSCAN clustering algorithm

A.3.1 Unsegmented sequence experiment

Sr No.	Type of Image	No of clusters	No of true nodules	Does cluster centroid coincide with ground truth
0	Negative	1	0	No
1	Negative	0	0	Yes
2	Negative	1	0	No
3	Negative	1	0	No
4	Negative	1	0	No
5	Negative	1	0	No
6	Negative	1	0	No
7	Negative	1	0	No
8	Negative	1	0	No
9	Positive	1	1	Yes
10	Negative	3	0	No
11	Negative	0	0	Yes
12	Negative	1	0	No
13	Negative	2	0	No
14	Negative	2	0	No
15	Negative	1	0	No
16	Negative	0	0	Yes
17	Negative	1	0	No
18	Negative	1	0	No
19	Negative	2	0	No
20	Positive	1	1	Yes
21	Positive	1	1	No
22	Positive	2	1	Yes
23	Positive	1	1	Yes
24	Positive	1	1	Yes
25	Positive	1	1	Yes
26	Positive	1	1	Yes
27	Positive	2	1	Yes
28	Positive	2	1	Yes
29	Positive	2	2	Yes,Yes
30	Positive	2	1	Yes
31	Positive	3	1	Yes
32	Positive	1	1	Yes
33	Positive	2	1	Yes
34	Positive	2	2	Yes,No
35	Positive	1	1	Yes
36	Positive	2	1	No
37	Positive	2	1	Yes
38	Positive	1	1	Yes
39	Positive	2	1	Yes
40	Positive	3	1	Yes
		57	24	24-Yes / 19 -No

FIGURE A.10: Aggregation results for unsegmented sequence experiment

A.3.2 Segmented sequence experiment

Sr No.	Type of Image	No of clusters	No of true nodules	Does cluster centroid coincide with ground truth
0	Negative	0	0	Yes
1	Negative	1	0	No
2	Negative	1	0	No
3	Negative	0	0	Yes
4	Negative	1	0	No
5	Negative	0	0	Yes
6	Negative	1	0	No
7	Negative	0	0	Yes
8	Negative	1	0	No
9	Positive	0	1	No
10	Negative	0	0	Yes
11	Negative	0	0	Yes
12	Negative	1	0	No
13	Negative	2	0	No
14	Negative	1	0	No
15	Negative	0	0	Yes
16	Negative	0	0	Yes
17	Negative	0	0	Yes
18	Negative	1	0	No
19	Negative	1	0	No
20	Positive	1	1	No
21	Positive	3	1	No
22	Positive	1	1	Yes
23	Positive	1	1	No
24	Positive	2	1	Yes
25	Positive	1	1	Yes
26	Positive	1	1	Yes
27	Positive	0	1	No
28	Positive	1	1	Yes
29	Positive	2	2	Yes,Yes
30	Positive	0	1	No
31	Positive	0	1	No
32	Positive	1	1	Yes
33	Positive	0	1	No
34	Positive	1	2	Yes,No
35	Positive	0	1	No
36	Positive	1	1	Yes
37	Positive	1	1	Yes
38	Positive	1	1	Yes
39	Positive	0	1	Yes
		29	23	22- Yes / 20- No

FIGURE A.11: Aggregation results for segmented sequence experiment

A.3.3 Segmented random experiment

Sr No.	Type of Image	No of clusters	No of true nodules	Does cluster centroid coincide with ground truth
0	Negative	0	0	Yes
1	Negative	1	0	No
2	Negative	1	0	No
3	Negative	1	0	No
4	Negative	1	0	No
5	Negative	1	0	No
6	Negative	0	0	Yes
7	Negative	1	0	No
8	Negative	1	0	No
9	Positive	1	1	No
10	Negative	0	0	Yes
11	Negative	1	0	No
12	Negative	0	0	Yes
13	Negative	1	0	No
14	Negative	0	0	Yes
15	Negative	1	0	No
16	Negative	0	0	Yes
17	Negative	1	0	No
18	Negative	0	0	Yes
19	Negative	1	0	No
20	Positive	2	1	Yes
21	Positive	2	1	No
22	Positive	1	1	Yes
23	Positive	3	1	No
24	Positive	2	1	Yes
25	Positive	1	1	No
26	Positive	1	1	Yes
27	Positive	1	1	No
28	Positive	1	1	No
29	Positive	2	2	Yes,No
30	Positive	1	1	No
31	Positive	1	1	No
32	Positive	1	1	Yes
33	Positive	0	1	No
34	Positive	2	2	Yes,No
35	Positive	1	1	No
36	Positive	1	1	Yes
37	Positive	1	1	No
38	Positive	1	1	No
39	Positive	2	1	No
		40	23	15- Yes / 27-No

FIGURE A.12: Aggregation results for segmented random experiment

A.3.4 Segmented sequence experiment with experts

Sr No.	Type of Image	No of clusters	No of true nodules	Does cluster centroid coincide with ground truth
0	Negative	1	0	No
1	Negative	3	0	No
2	Negative	1	0	No
3	Negative	2	0	No
4	Negative	1	0	No
5	Negative	0	0	Yes
6	Negative	1	0	No
7	Negative	2	0	No
8	Negative	1	0	No
9	Positive	1	1	Yes
10	Negative	0	0	Yes
11	Negative	1	0	No
12	Negative	1	0	No
13	Negative	1	0	No
14	Negative	1	0	No
15	Negative	1	0	No
16	Negative	1	0	No
17	Negative	2	0	No
18	Negative	0	0	Yes
19	Negative	1	0	No
20	Positive	1	1	Yes
21	Positive	2	1	Yes
22	Positive	1	1	Yes
23	Positive	2	1	Yes
24	Positive	1	1	Yes
25	Positive	1	1	Yes
26	Positive	1	1	Yes
27	Positive	2	1	Yes
28	Positive	1	1	Yes
29	Positive	2	2	Yes,Yes
30	Positive	2	1	Yes
31	Positive	1	1	Yes
32	Positive	1	1	Yes
33	Positive	1	1	Yes
34	Positive	2	2	Yes,No
35	Positive	1	1	Yes
36	Positive	2	1	Yes
37	Positive	1	1	Yes
38	Positive	1	1	Yes
39	Positive	1	1	Yes
		49	23	25- Yes / 17 -No

FIGURE A.13: Aggregation results for segmented sequence experiment using experts

Bibliography

[1] Nadeem S. Park J. H. Baker K. Kaufman A. 2. Boorboor S. "Crowdsourcing lung nodules detection and annotation. " In: *Medical Imaging 2018: Imaging Informatics for Healthcare, Research, and Applications.* (2018). DOI: 10.1117/12.2292563.

[2] S. G. Armato et al. "The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans". In: *Medical physics, 38(2):* (2011), 915–931.

[3] Macmahon H. Goo J. M. Rubin G. D. Schaefer-Prokop C. M. Naidich D. P. Bankier A. A. "Recommendations for Measuring Pulmonary Nodules at CT: A Statement from the Fleischner Society". In: *Radiology, 285(2)* (2017), 584–600.

[4] Ferlay J. Soerjomataram I. Siegel R. L. Torre L. A.-Jemal A. Bray F. "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries". In: *CA: A Cancer Journal for Clinicians, 68(6)* (2018), 394–424. DOI: 10.3322/caac.21492.

[5] Perez-Rovira A. Kuo W. Tiddens H. A. W. M.-Bruijne M. D. Cheplygina V. "Early Experiences with Crowdsourcing Airway Annotations in Chest CT." In: *Deep Learning and Data Labeling for Medical Applications Lecture Notes in Computer Science* (2016), 209–218. DOI: 10.1007/978-3-319-46976-8_22.

[6] Balachandar-N. Chon A. "Deep Convolutional Neural Networks for Lung Cancer Detection. " In: (2017).

[7] K. Q. Weinberger G. Huang Z. Liu and L. van der Maaten. "Densely connected convolutional networks". In: *CVPR* (2017).

[8] D. R. Kim H. Tang and X. Xie. "Automated pulmonary nodule detection using 3D deep convolutional neural networks ". In: *in Proc. IEEE 15th Int. Symp. Biomed. Imag. (ISBI), Washington, DC, USA* (Apr. 2018), 523–526.

[9] Najeeb S. Shahriyar A. Abdullah Z. R. Haque M. A. Hossain S. "A Pipeline for Lung Tumor Detection and Segmentation from CT Scans Using Dilated Convolutional Neural Networks. " In: *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2019). DOI: 10.1109/icassp.2019.8683802.

[10] Jeff Howe. "The Rise of Crowdsourcing ". In: (2006). URL: https://www.wired.com/2006/06/crowds/.

[11] S. Ren K. He X. Zhang and J. Sun. "Deep residual learning for image recognition". In: *CVPR* (2016).

[12] Russakovsky O. Fei-Fei L. Grauman K. Kovashka A. "Crowdsourcing in Computer Vision ". In: *Foundations and Trends® in Computer Graphics and Vision, 10(3)* (2016), 177–243. DOI: 10.1561/0600000071.

[13] Xie X. Shen-L. Liu S. Li Y. "Reverse active learning based atrous DenseNet for pathological image classification ". In: (2019).

[14] Kooi T Bejnordi-B. E Setio A. A. A Ciompi-F Ghafoorian M van der Laak J. A Van Ginneken B Litjens G and C. I. Sánchez. "A survey on deep learning in medical image analysis". In: *Medical image analysis 42* (2017), 60–88.

[15] "LUNA16 - Grand Challenge. (n.d.)." In: (). URL: https://luna16.grand-challenge.org/.

[16] Prasad P. Alsadoon-A. Singh A. Makaju S. and A. Elchouemi. "Lung Cancer Detection using CT Scan Images." In: *Procedia Computer Science* 125 (2018), pp. 107–114.

[17] Jorg S. E Xiaowei X Martin E Hans-Peter K. "A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise". In: *KDD'96 Proceedings of the Second International Conference on Knowledge Discovery and Data Mining* (1994), pp. 226–231.

[18] "Medical image computing and computer-assisted intervention". In: *Springer* (2013).

[19] Ulas Bagci Naji Khosravan. "S4ND: Single-Shot Single-Scale Lung Nodule Detection. " In: *MICCAI (2) 2018* (2018), pp. 794–802.

[20] P. Fischer O. Ronneberger and T. Brox. "U-net: Convolutional networks for biomedical image segmentation ". In: *MICCAI* (2015).

[21] Murchison J. T. Beek E. J. R. V. Goatman-K. A. O'Neil A. Q. "Crowdsourcing Labels for Pathological Patterns in CT Lung Scans: Can Non-experts Contribute Expert-Quality Ground Truth?" In: *Lecture Notes in Computer Science Intravascular Imaging and Computer Assisted Stenting, and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis* (2017), 96–105. DOI: 10.1007/978-3-319-67534-3_11.

[22] S. Uzelaltinbulat and B. Ugur. "Lung tumor segmentation algorithm." In: *Procedia Computer Science* 120 (2017), pp. 140–147.

[23] M. Winkels and T. S. Cohen. "3D G-CNNs for pulmonary nodule detection ". In: *arXiv:1804.04656. [Online]* (2018). URL: https://arxiv.org/abs/1804.04656.

[24] H. Xiao X. Jin S. Yan Y. Chen J. Li and J. Feng. "Dual path networks". In: *NIPS* (2017).

[25] Liu C. Fan W. Xie X. Zhu W. "DeepLung: Deep 3D Dual Path Nets for Automated Pulmonary Nodule Detection and Classification ". In: *IEEE WACV* (2018).

[26] Doyle A. Hilten A.V. Hirth M. Inel O. Madan C.-Mavridis P. Spiers H. Cheplygina v. Ørting S. "A survey of crowdsourcing in medical image analysis ". In: (2019). URL: https://arxiv.org/pdf/1902.09159.pdf.