An Ultrasound Video-based AI-assisted Point-of-Care Tool for Tuberculosis Screening

Mukul Ranjan*
Wadhwani Institute for Artificial Intelligence
Mumbai, India
mukul.ranjan@wadhwaniai.org

Amrit Saha
Wadhwani Institute for Artificial Intelligence
Mumbai, India
amrit.saha@wadhwaniai.org

Ankit Bhardwaj
Wadhwani Institute for Artificial Intelligence
Mumbai, India
bhardwaj@wadhwaniai.org

Vishal Agarwal
Wadhwani Institute for Artificial Intelligence
Mumbai, India
vishal@wadhwaniai.org

Alpan Raval
Wadhwani Institute for Artificial Intelligence
Mumbai, India
alpan@wadhwaniai.org

Rahul Panicker
Wadhwani Institute for Artificial Intelligence
Mumbai, India
rahul@wadhwaniai.org

Abstract

The current screening pipeline for Tuberculosis(TB) screening has many challenges including the requirement of huge infrastructure for X-ray testing, delay due to different tests and drop-off due to the complicated process. It is also not point-of-care and requires involvement of experts. In this work, we propose to build a much safer and portable ultrasound-based point-of-care screening tool using AI. Our proposed AI method not only surpasses radiologists in classifying the ultrasound-videos with a large margin but also provides a proper explanation to its outcomes.

1 Introduction

Tuberculosis(TB) is the top cause of death from infectious disease globally, causing an estimated death of 1.7 million among total of 10.4 million cases in 2016 out of which India alone accounted for more than 27% of the cases[1]. Currently, there are several tests for screening and diagnostics of TB with various sensitivity ans specificity values including chest x-ray, smear microscopy and rapid molecular diagnostic tests like CBNAAT but there are different challenges associated with them.

Chest x-ray is assumed to be the best screening tool with sensitivity of 90% at 60% specificity, but if a chest x-ray shows scar to the lungs there can be several possibilities, including active TB, healed TB or some other conditions, hence this is useful but not conclusive. Also, presumptive TB patients(patient who presents with symptoms or signs suggestive of TB) are given free vouchers to go for X-Ray testing but there may be significant delay and drop-off in between whose rate varies from 4% to 40% depending on geographic and socio-economic factors. Smear microscopy, another test used for screening, is not sufficiently sensitive as it needs a high concentration of bacilli to show a positive result. Rapid molecular diagnostic tests like CBNAAT can detect TB bacilli even when they are present in low concentration. However, the availability and access to these tests is not universal.

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

Preprint. Under review.

Requirement of huge infrastructure is one of the major challenges in the current screening tools. Due to this, these techniques are not point-of-care(POC). A portable screening tool based on ultrasound imaging (which is much safer than x-ray) embedded with an AI system will be able to solve a lot of challenges associated with the current screening methods. With a portable ultrasound device we will have POC solution resulting in no drop off. Due to its POC nature, this tool will allow a health visitor to screen for TB without any support from an expert.

Our deep learning[2] based approach explained in this paper not only surpassed radiologists in the classification of ultrasound videos but also provided proper explanations to its outcomes. We first formulated our problem as a binary image classification problem where each frame is assigned a 0-1 label based on the presence or absence of abnormalities in it. Then we used aggregation method to predict the label of the video. Our contributions can be summarized as follows:

- We are the first to use deep learning based approach on ultrasound imaging of Tuberculosis.
- We used ResNet101[3] for the binary classification of the frames of the ultrasound videos and after after using aggregation method for classifying videos, our model surpassed radiologists on both AP and AUC of ROC curve metric at suitable threshold(0.1).
- We proposed to use SSD[4] and CFSSD framework for the frame level detection of the abnormal features. Our method provides good explainability to the binary classification model by detecting abnormal features in the positive videos with a recall of 74.34% at the precision of 83.44% on val set.
- We did detailed experimentation for abnormal feature detection with various kind of labels including detection of abnormal features as a single positive class (we call it 1-class detection), detection of distinct abnormalities or positive features (5-class detection), detection of top 2 prevalent positive feature with almost no class imbalance (2-class detection).

2 Related Works

In the TB space, to the best of our knowledge there have been no work done which uses ultrasound imaging. Various formulation regarding detection and classification of chest x-ray have been done using deep learning. Qin, et al.[5] conducted a retrospective evaluation of three DL systems namely CAD4TB, Lunit INSIGHT, and qXR for detecting TB-associated abnormalities in chest x-ray. They found that the area under the roc-curve of the three system was similar.Lakhani, et al.[6] used an ensemble of the AlexNet[7] and GoogLeNet[8] to get an AUC of 0.99. Heo, et al.[9] proposed using demographic features along with VGG19[10]. They found that using demographic features increased the AUC value by 0.0138 (0.9075 to 0.9213) in the test set.

Looking in the ultrasound domain we could not find any work which has used ultrasound imaging for TB classification or detection. There have been some works which uses ultrasound imaging for thyroid cancer[11, 12, 13] and breast lesion detection and classification[14].

3 Datasets

Datasets are the most important aspect of of any deep learning research. We collected a dataset for 46 subjects out of which 28 were diagnosed TB-positive through other tests while 18 were TB-negative.

3.1 Protocol for Data Collection

For collecting data on these subjects, the radiologist had with them, the x-ray and other test results of the subject. Using the x-ray, the radiologist was able to localize on the region of the chest that would likely result in visible features in the ultrasound image. One point to note here is that for negative subjects, the x-rays were not useful in localizing as they did not show any abnormal features. Thus, their data was collected randomly.

We have 30 second long videos, usually multiple for each subject. In the case of TB-positive patients, the radiologist recorded the videos after visually confirming the presence of relevant features. Thus, the videos that we have for TB-positive patients are extremely demonstrative and the distribution of data is likely to be very different from what our proposed use case is. In the case of negative patients, the videos are fairly random.

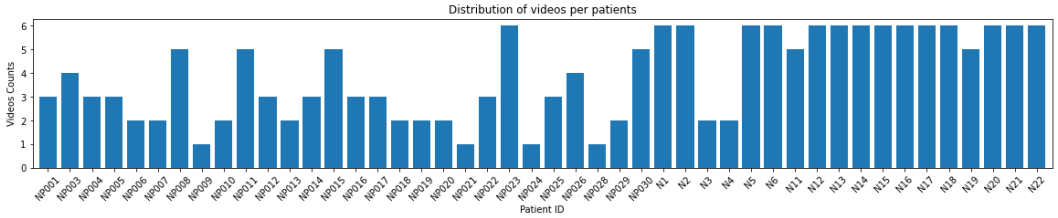

Figure 1: Distribution of videos for all patients

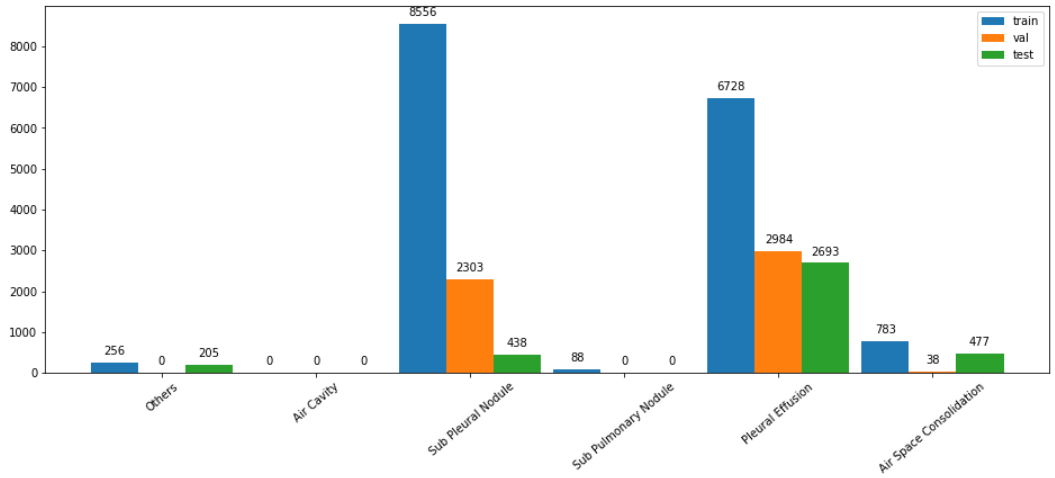

Figure 2: Distribution of features in different splits

3.2 Data Distribution

For each of the 46 subjects we have variable number of videos ranging from 1 to 6. Positive patients are identified by the prefix **NP** while negative subjects are identified by prefix **N**. The distribution of the number of videos for positive and negative patients can be seen in bar graph shown in figure[1].

In total we have collected 179 videos out of which 81 are positive label videos(from positive subjects) and 98 are negative label videos (from negative subjects).

These videos have varying number of frames bounded by 1000. The frame rate is 33 frames per second and a linear probe of 10 MHz is used for taking videos.

Each of these videos are annotated at frame level by two annotators. For each frame with abnormal features, features are identified as classified into six different categories(assigned label from 0 to 5) as shown in the table[2].

Looking at the frames, we have in total 147,928 unique frames and 295,856 annotations. One of the major problems with our dataset is the high imbalance between features which can be seen in the feature[3]. There are two classes namely Pleural Effusion and Sub Pleural Nodule with almost equal distribution. Another thing to note here is that there is no frame for class Air Cavity and hence we removed this class from our experiment.

We created the train, test and val splits based on patients and not based on frames. Since this is a video dataset, creating splits based on frames can cause spills between the splits as subsequent frames in the video of the same patients will be almost similar.

3.3 Difficulties with the Datasets

Major drawback of this split is that there is an imbalance between the presence of features in different splits as seen the figure[2]. We note that features **Others** which is present in train set and test set but not in the val set. Similarly feature **Air Space Consolidation** is present in train set but not in

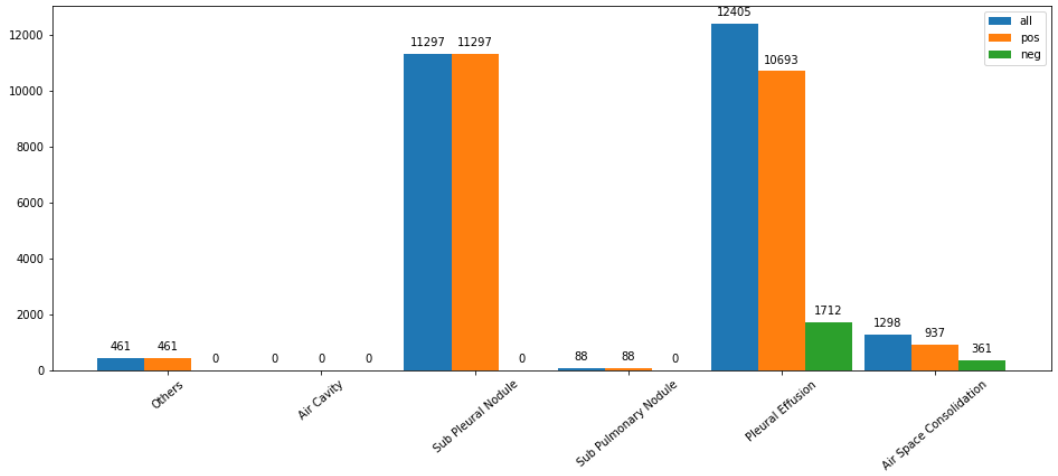

Figure 3: Distribution of features in frames

Table 1: Feature wise Disagreement

Feature Name	Total Frames	Total Boxes	Mean IOU	Median IOU	Match	HQ Match
Others	461	461	0	0	0	0
Air Cavity	0	0	0	0	0	0
Sub Pleural Nodule	6191	7161	0.3616	0.4432	4189	3079
Sub Pulmonary Nodule	79	79	0	0	0	0
Pleural Effusion	8683	8697	0.2637	0	3708	2762
Airspace Consolidation	1298	1298	0	0	0	0

the val and test set. Another problem which we found in the dataset is the agreement between the annotators. As compared to chest x-ray imaging, chest ultrasound images are really difficult to find features in. Feature agreement between two experts among all the positive frames is 63%. Here we define agreement in terms of **set_IOU(set Intersection Over Union)**, as the ratio of total number of frames where both annotators agrees for them to belongs to a class and the union of all the frames of that class.

For defining set_IOU we first define **set_Intersection** as the total number of frames where both annotator agrees to be from same class and **set_Union** as the union of the total frames with unique labels.

$$set\_IOU = \frac{set\_Intersection}{set\_Union} \tag{1}$$

We use another term, **Intersection Over Union(IOU)** or the **Jaccard Index** to define the bounding box agreement between the annotators. It is defined as the area of intersection between the box annotated by annotator 1 and annotator 2, divided by the area of union between the box annotated by annotator 1 and annotator 2.

$$IOU = J(A, B) = \frac{A \cap B}{A \cup B} \tag{2}$$

where A and B denote the bounding box annotated by annotator 1 and annotator 2, respectively.

The agreement for different classes can be seen in table[1]. It can be noted that mean IOU and median IOU of agreements are very less. **Match** is the number of frames on which both the annotators agree it to be of a specific class. **HQ Match** is a subset of Match that contains all the images with IOU between the annotators greater than or equal to 0.5. Another difficulty which we found associated with this dataset is the similarity between Sub Pleural Nodule and Pleural Effusion. As seen from the table[1], it is clear that both annotators are agreeing only in 42% cases for Pleural Effusion. This looks very clear when we look at figure[4]. It is really difficult for human eye to distinguish between Pleural Effusion and Sub Pleural Nodule in some cases.

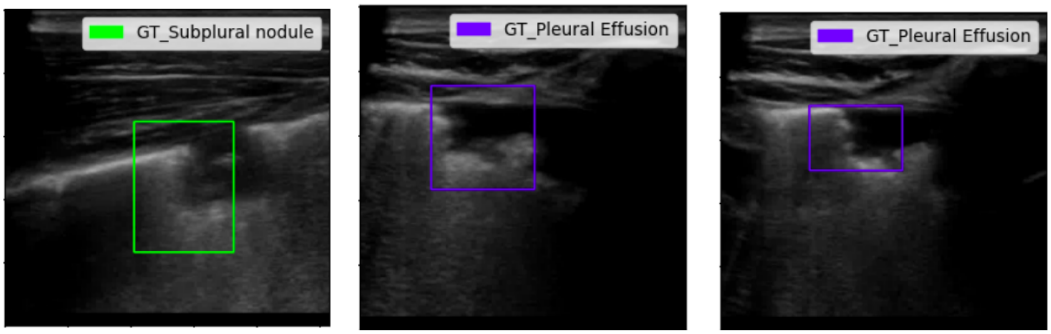

Figure 4: Feature Anomalies in two major classes

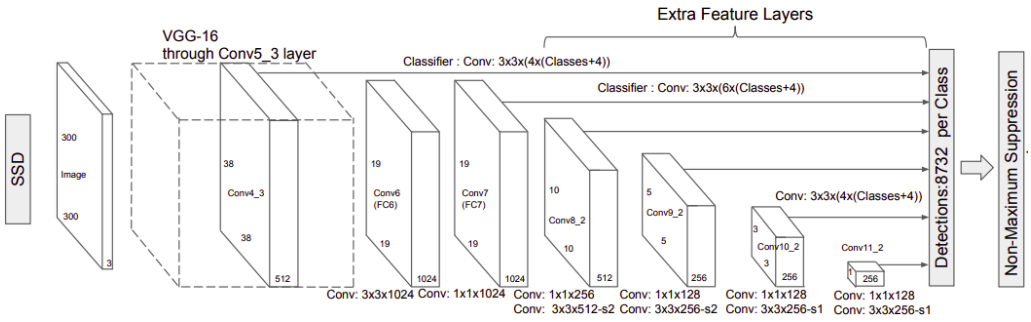

Figure 5: SSD Architecture with VGG16 backbone(Source: Liu, et al.[4])

4 Model Architecture

Our initial experiment was the Binary classification of video frame. For this we experimented with Squeezenet, Inceptionv3, DenseNet, ResNet50, and ResNet101. The architecture of our best performing model ResNet101. For our frame level detection experiments we used SSD[4] with various ResNet[3] and VGG16[10] as backbone. The architecture with VGG16 backbone as developed by Liu, et al. is shown in figure[5]. We also used a multi-task learning framework CFSSD(Dalmia, et al.) in order to incorporate classification and detection branch in a single backbone. The architecture of CFSSD contains contains SSD with a separate classifier branch. That is shown in figure[6]

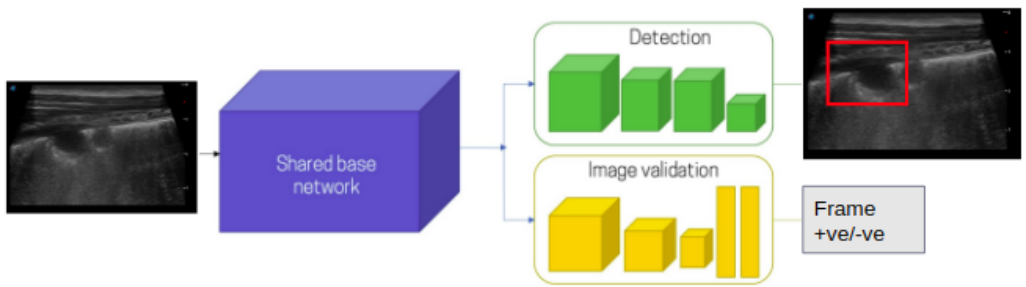

Figure 6: CFSSD Architecture, Source : Dalmia et al.[15]

Table 2: Feature No and feature Type

Feature No.	Feature Name
0	Others
1	Air Cavity
2	Sub Pleural Nodule
3	Sub Pulmonary Nodule
4	Pleural Effusion
5	Airspace Consolidation

5 Experiments

We performed various sets of experiments throughout this project. Beginning from binary classification to detection on frames for different classes, we can first summarize our experiments as follows:

- Binary classification of the video frames using transfer learning from imagenet and upsampling positive frames to have a prevalence of 50%.
- 5-class multi task learning of frame level detection and classification using CFSSD framework.
- 1-class multi task learning of frame level detection and classification using CFSSD framework.
- 1-class multi task learning of frame level detection and classification using CFSSD framework only on a single annotator's annotation.
- 2-class frame level detection using SSD framework.
- 2-class frame level detection using SSD framework only on high quality annotations
- 1-class frame level detection using SSD framework.

5.1 Evaluation Pipeline and Metrics

Since we are using a Multi-task learning framework, it is really important to understand various evaluation pipelines which are being used in our work. For the model with both classification and detection branch in it, we have three evaluation pipeline since three different task is being done by the framework.

- **Object Detection pipeline**
- **Frame Classification Pipeline**
- **Video Classification Pipeline**

Detection Evaluation Pipeline is shown in the figure[7]. For a given input image, the detection branch gives us encoded boxes along with the classification score for each boxes detected. These predicted boxes are decoded to remove background and for Non-Max-Suppression(**NMS**,the process of removing multiple boxes with high (greater than the **NMS threshold**) Intersection Over Union(IOU) between them). After decoding, we get our final prediction as the bounding box coordinates and the class which the predicted boxes belongs to, along with its classification score. To calculate various metrics, we used the methods discussed in [16], according to which detections are assigned to ground truth objects and judged to be true/false positives by measuring bounding box overlap. According to [16], to be considered as a correct detection, the **overlap ratio** a_o(we denote this overlap ratio as **iou**) between the predicted bounding box B_p and ground truth bounding box B_{gt} must exceed 0.5 (50%) by the equation (3)

$$iou = \frac{B_p \cap B_{gt}}{B_p \cup B_{gt}} \tag{3}$$

Overlap Ratio of 50% is not a good value for us. Looking at very low agreement between the annotators and also from table (1) we can see that **mean IOU** between the annotators itself is less than 50%, it is reasonable to consider the **overlap ratio** or **iou** threshold to be equal to **0.3**(30%).

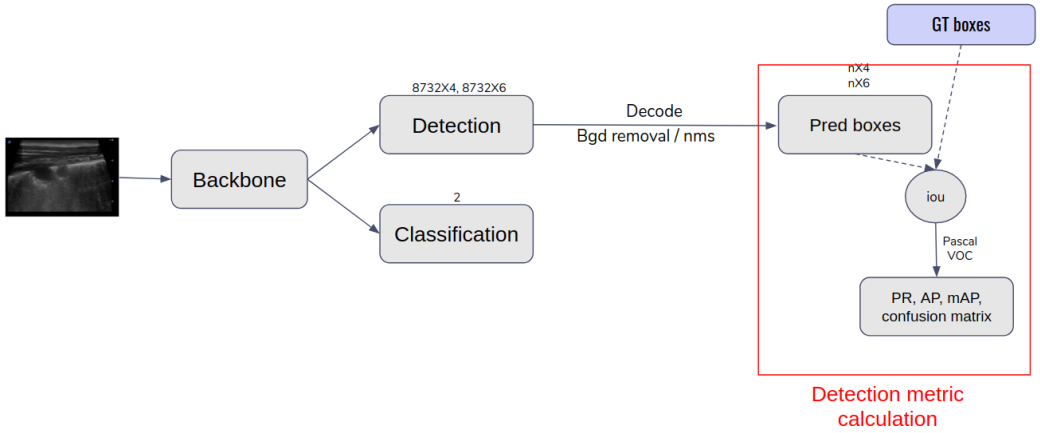

Figure 7: Object Detection Pipeline

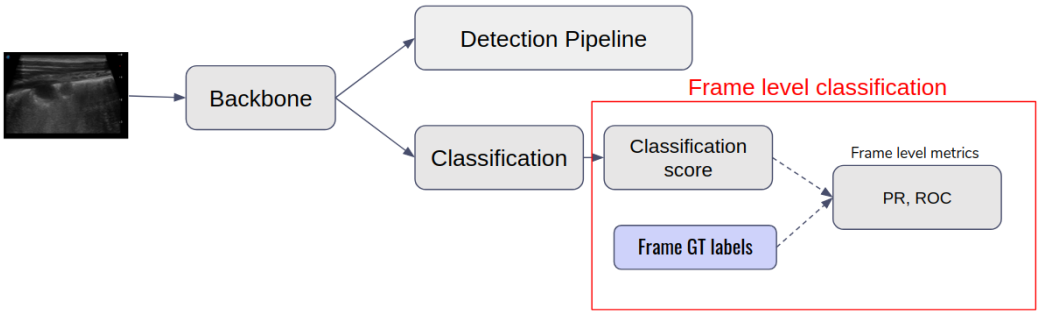

Figure 8: Frame Level Classification Pipeline

After getting the detected labels for each class as True/False positive we calculate the precision and recall for each class which is given by equation (4).

$$P = Precision = \frac{TP}{TP + FP}, \; Recall = R = \frac{TP}{TP + FN} \tag{4}$$

After getting the precision and recall we calculated **F1 Score** by taking their harmonic mean. Please note here that the F1 Score is the **primary metric** of frame level object detection.

COCO MAP, another metric which we used to check the performance of our detector in terms of standard object detection metric. It is the almost similar to that used in COCO[17] object detection challenge. **MAP** is the mean of the Average precision of all the classes. But the primary challenge metric of the COCO[17] object detection challenge, takes the mean of MAP for a range of **overlap ratio** or **iou**, from 0.5 to 0.95 at an internal of 0.05. We used similar metric here as well with a slight modification in the range of **iou**, which we used from 0.3 to 0.6 with an interval of 0.03.

Frame Classification Pipeline is shown in figure[8]. The common backbone is extended with an Adaptive Average Pooling layer, which is later flattened and extended further with a ReLu, Dropout and Linear layer to give a final output. We further use a softmax layer to get the **classification score** for each frame. We use this predicted score to calculate frame classification metrics. The **primary metric** for the evaluation of frame classification is **ROC-AUC**. We also calculated the confusion matrix and the PR-Curve(Precision-Recall Curve) to measure the performance of our model.

Video Classification Pipeline is shown in figure[9] and in figure[10]. After getting the frame-level classification score, we can classify each frame as predicted positive or predicted negative, for a given threshold(we call this **frame_score_thres** or **p1**). We define video score by the method of

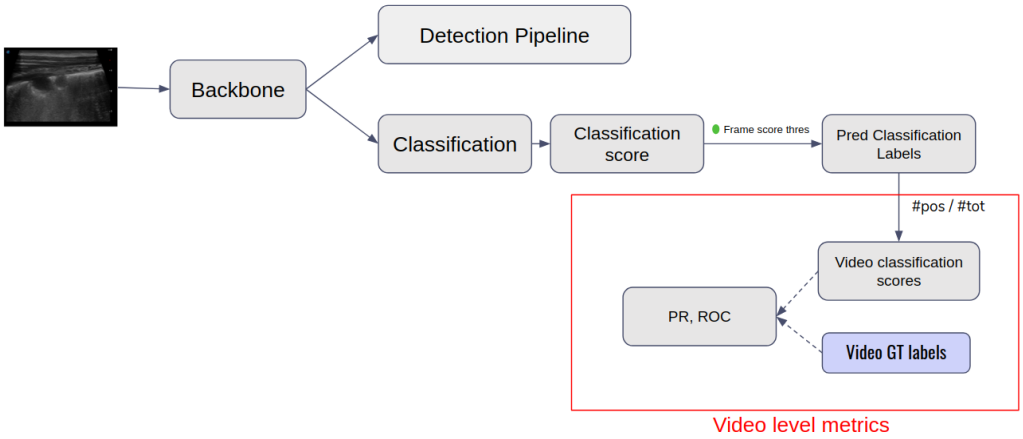

Figure 9: Video Level Classification Pipeline

aggregation i.e for a given **p1** we count the total number of predicted positive box and define **vide score** by equation (5). We define another threshold **p2** to get the video classification result. By varying $p1$ and $p2$ we get different **PR curves(Precision-Recall Curve)** and **ROC curves (receiver operating characteristic curve)** for video classification results.

$$Video\ Score = \frac{\#positive\ frames}{\#total\ frames} \tag{5}$$

Optimization of **p1** is done to maximize the value of ROC-AUC for video classification since our goal is the classification of videos.

Our **primary metric** for the frame level detection is **F1-Score** and for frame and video classification it is **ROC-AUC**.

5.2 Binary Classification with Transfer Learning using ImageNet weight

We first modeled our problem as binary frame classification problem where we fine-tuned various state of the art models on our dataset. We experimented with Squeezenet[18], Inceptionv3[8], DenseNet[19], ResNet50[3], and ResNet101[3]. Based on the results from validation set, we found out resnet101[3] to be the best performing model. We optimized our model on validation set at video level **ROC-AUC** metric as explained in previous section. A simple pipeline with methodology for frame and video classification can be seen in the figure[10].

Since we are the first to use deep learning based methods to solve the problem of Tuberculosis using ultrasound imagery, we could not find any baseline other than the radiologists labels. Due to this reason, we considered this method as our another baseline for multi-task learning of combined classification and detection problem (CFSSD[15]). In our plot which we showed in the later part of the paper we denoted this method with short term **clf_resnet101**. While we showed the PR-curve(Precision-Recall Curve) and ROC-curve(Receiver Operating Characteristic curve) in the later part of the curve, the confusion matrix for train and test set at optimal **p1** of 0.1 can be seen in figure (11).

5.3 5-class detection and classification using CFSSD

In this experiment we did object detection of the the five features which we had in our datasets(as seen from figure [3], there are no frames for class Air Cavity, therefore we removed that class from our experiments). The common backbone network for the best experiment of this model is **resnet34**. When we look at the table (1), we find that there is no match for class Others, Sub Pulmonary Nodule and Air Space Consolidation. This implies that it is really difficult to identify those features from the USG frames for radiologists. This experiment shows that our model also could not detect any of those features. The confusion matrix for the frame level feature detection part of this experiment is shown in figure (13).

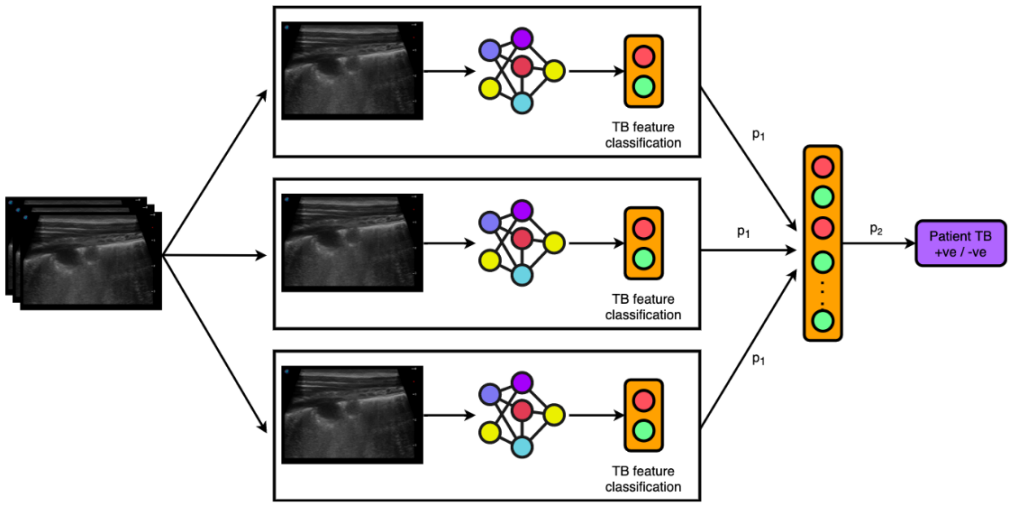

Figure 10: A Simple Pipeline with methodology for frame and video classification: Our framework take a video as an input and for each frame get a classification score **p1**, which is fine-tuned keeping **p2** constant to optimize the video classification result.

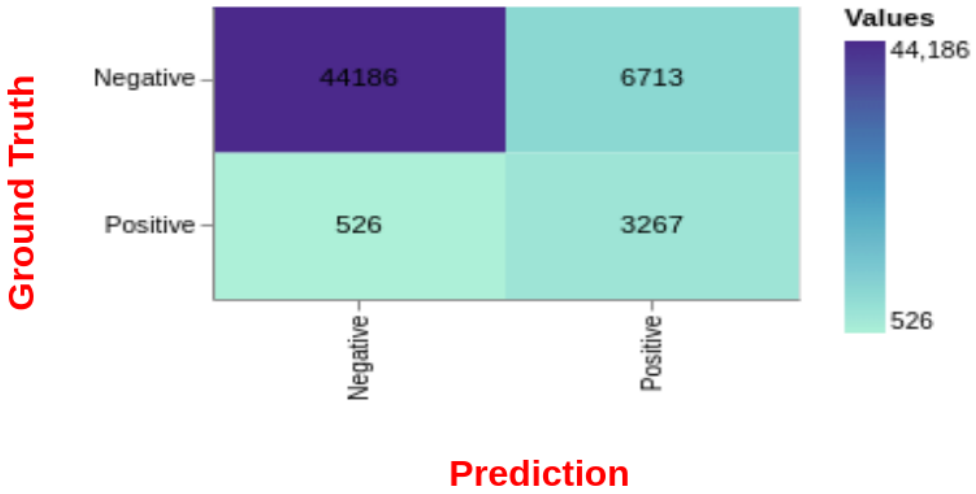

Figure 11: **Frame Classification confusion Matrix for clf_resnet101 on Test Set**: We see that even though this model is able to predict a most of the positive features to the positive class and most of the negative frames are getting predicted to the negative class, positive class is getting a large amount of its prediction from negative class, which does not seems to be good when we look at one glance, but important things to remember here is that we are not optimizing our frame level metrics but the video level metric, and that is geeting optimized at **p1 = 0.1** as seen from the figure (12)

.

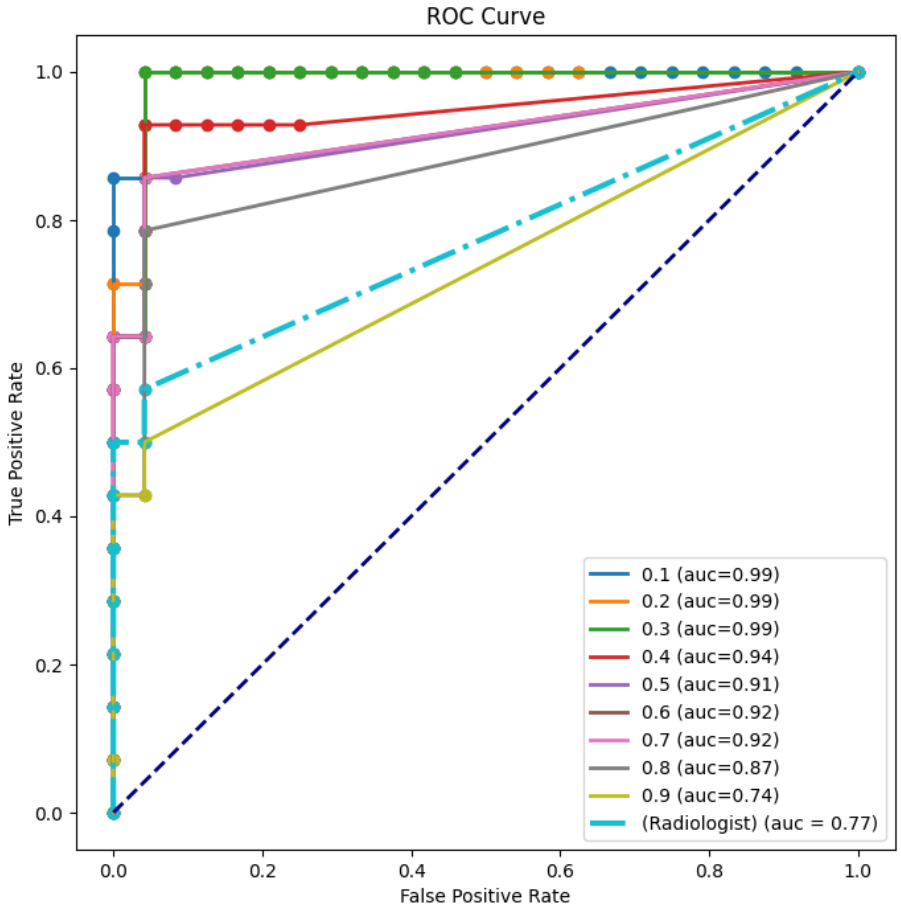

Figure 12: **Video level ROC curve for clf_resnet101 at different value of p1 on val set**: We optimized the value of p1 at val set, to get the best value of video level ROC-AUC. The best value of p1 which we got from the curve is 0.1.

The detection branch of CFSSD[15]/SSD[4] model gives a large number of boxes. The process of decoding which we are using can be seen in the figure (14). The **conf score** shown in the figure is the threshold to filter out the boxes for each class numbered from 1 to n(for our case n = 5). After this NMS is applied to each filtered boxes to get the final prediction boxes along with their label and box classification score.

Optimization of Confidence Score is one of the major problems for our task. Optimizing **MAP**, will always result in the lowest threshold and that is not our concern as well. We want our model to be well explainable, i.e. **If the classification branch predicts a frame to be positive then detection branch should be able to predict the correct bounding box in the frame**. But optimizing only **Recall** will also result in getting the lowest possible score as confidence score, but lower the confidence score higher the recall and lower the precision(since false positive will increase at lower confidence score). For this reason we chose to use **F1 Score** as the primary matrix for detection. Figure (15) shows the variation of F1 score for Sub Pleural Nodule and Pleural Effusion class and since for other classes we always got F1 score to be 0, we didn't add that in the plot.

.

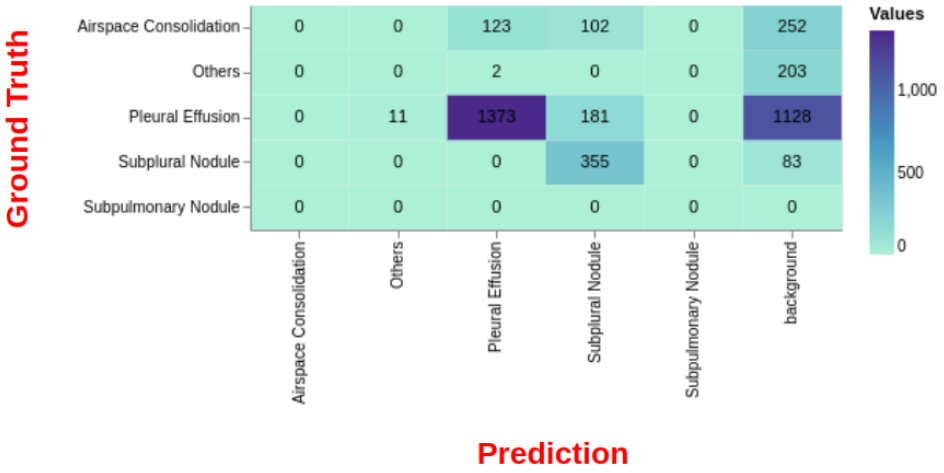

Figure 13: **Confusion Matrix for 5-class detection using CFSSD on Test Set**: Figure shows the confusion matrix for 5-class Detection model using CFSSD. Important thing to note here is that our model also could not predict any of the feature from class Others, Sub Pulmonary Nodule and Air Space Consolidation.

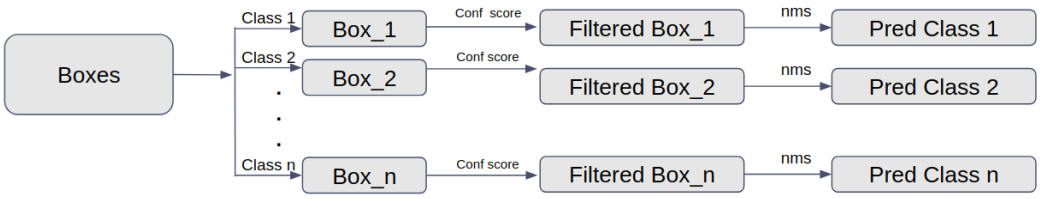

Figure 14: **Decoding in Detection branch of SSD[4]/CFSSD[15]**: For each frame, all the boxes for all the classes(numbered from 1 to n, for 5-class CFSSD n = 5) is filtered out using a confidence threshold(we call this **Confidence Score**). After which Non-Maximum-Suppression(NMS) is applied to get the final prediction boxes

.

Results from the classification branch of this 5-class CFSSD model appears to be as good as good as the clf_resnet101 model. The PR curve and the ROC curve can be seen in the figure (16).

Even though model seems to be doing slightly worse than the baseline(clf_resnet101), in terms of the frame level performance, when we look at the video level performance of CFSSD, we find it to be doing far better than the baseline on test set. Figure (17) compares the video level performance of 5-class CFSSD and two of the baseline radiologist and clf_resnet101.

5.4 1-class detection and classification using CFSSD

The quality of dataset is not good when we look each features out of all the five separately, but what if we assign all the five features a single label-**positive**? In this case, we saw that the agreement between both annotator is 63%, which seems to be much better as compared to before. Motivated by this intuition we modelled all of the five positive features as a single positive class - and did our object detection using CFSSD framework as discussed in section 4. The idea seems to be really good and improved the result of the detection branch, but the result on frame and video classification part became much worse the baseline clf_resnet101. Similar to the 5-class CFSSD part, the **Confidence score** is optimized here as well in the similar way. We considered the threshold for which F1 score was maximum on val set. The optimal value of **Confidence score** as seen from figure (18) came out to be 0.5 for which we got the F1 score of 78%, with a recall of 74% precision of 84%. After getting the optimal threshold we evaluated our model at the test set and found that the F1-score drop to a

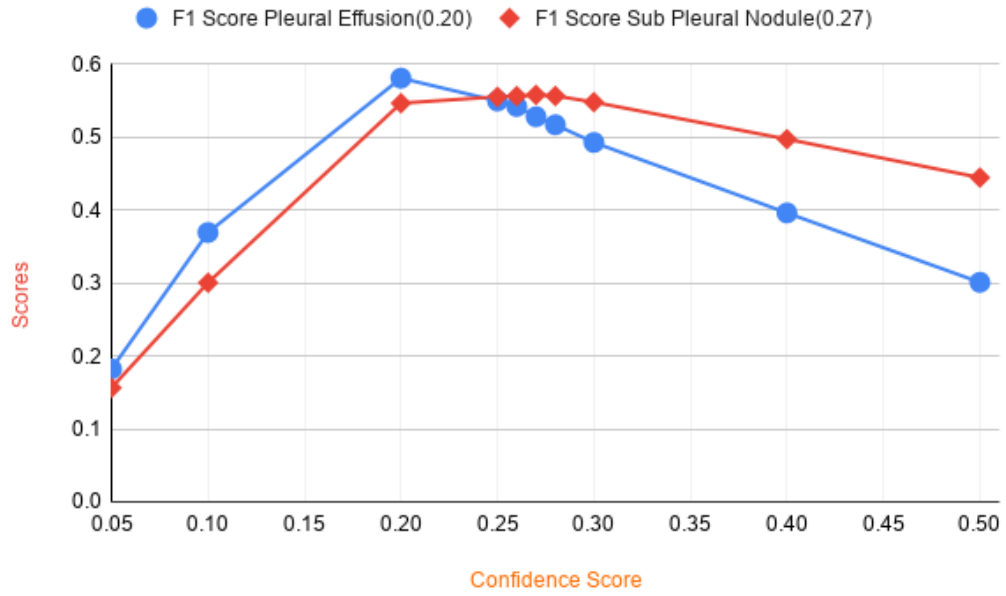

Figure 15: **Effect of thresholds on F1 Score for two major detected output in 5-class CFSSD**: Optimizing the F1 Score for each class individually. We found that optimal value of confidence score for class **Pleural Effusion** is 0.20 while for class **Sub Pleural Nodule** the optimal value is 0.27

.

value of 61.45%. This drop in the result was expected since all the outliers are deliberately kept in the test set.

Results from the classification branch for this experiment seems to be worse that 5-class CFSSD model. The PR and ROC curve for frame classification can be seen in the figure(19) while for video classification it can be seen in the figure (20).

5.5 1-class CFSSD and only on single annotator's annotation

Even after considering all five positive features as a single positive class, we still have bad data points with us because, for each image frame we are use two slightly different annotations. Our intuition here was that maybe using two different annotations for a single frame is hindering the model's performance instead of helping. Motivated by this idea we considered the annotations of a single annotator to train and evaluate our model.

In this experiment, detection head of the model was doing almost as good as the previous 1-class CFSSD experiment where we considered the annotations of both the annotators together. Looking at the classification branch of the model, this experiment did much better than the previous experiment, but still in terms of the performance of classification branch, 5-class CFSSD is the best among all the models.

The PR and ROC curves for frame classification with annotation from only one annotator in test set can be seen in the figure(22) while for video classification it can be seen in the figure (23). Looking at the figure (22), it seems that this model is doing better than the our baseline clf_resnet101, but the important thing to note here is that our baseline contains the frame labels from both annotators but this model consists of the annotations from only one annotator for each frame. When we compare the performance of models on the test set with labels from both annotator combined as done in the baseline clf_resnet101 performance seems to drop as seen in the figure (21).

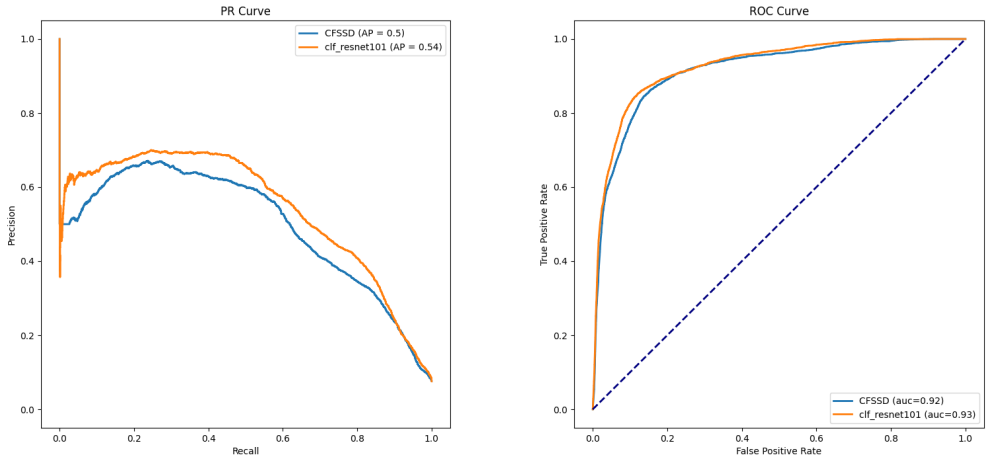

Figure 16: **PR and ROC curve for frame level classification of the 5-class CFSSD Model on Test Set**: The figure compared the result of multi-task learning framework CFSSD with resnet34 backbone and clf_resnet101(resnet101 based binary classification model)

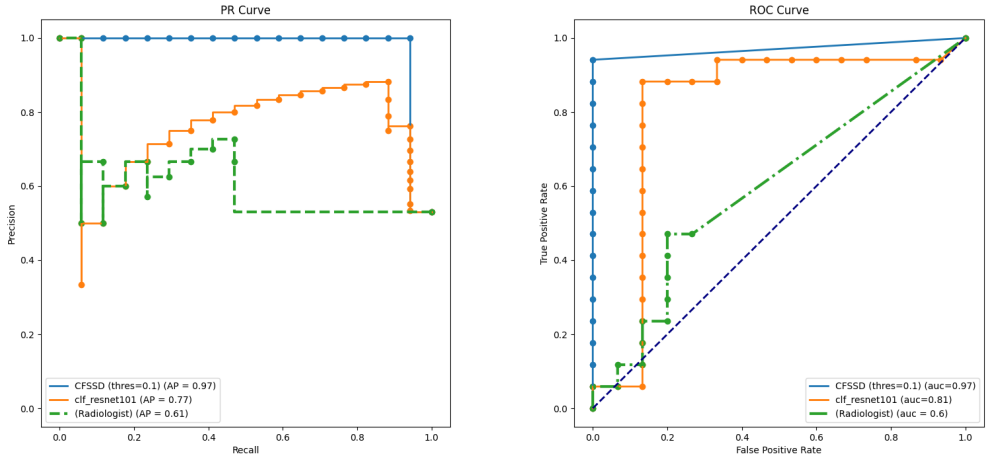

Figure 17: **PR and ROC curve for video level classification of the 5-class CFSSD Model on Test Set**: The figure compared the result of multi-task learning framework CFSSD with resnet34 backbone and clf_resnet101(resnet101 based binary classification model) and radiologist. CFSSD model seems to be doing far better than the both radiologists and clf_resnet101

5.6 2-class frame level detection using SSD

Other than the poor agreement among the annotators, another major problem which we dealt with while experimentation is the problem of class imbalance. In 1-class detection we don't have any class-problem since there is only single class, but as we increase the number of classes in our experiments this problem become more and more dominant. Since the number of samples in class Sub Pleural Nodule and Pleural Effusion is almost same, for this experiment we considered only these two classes.

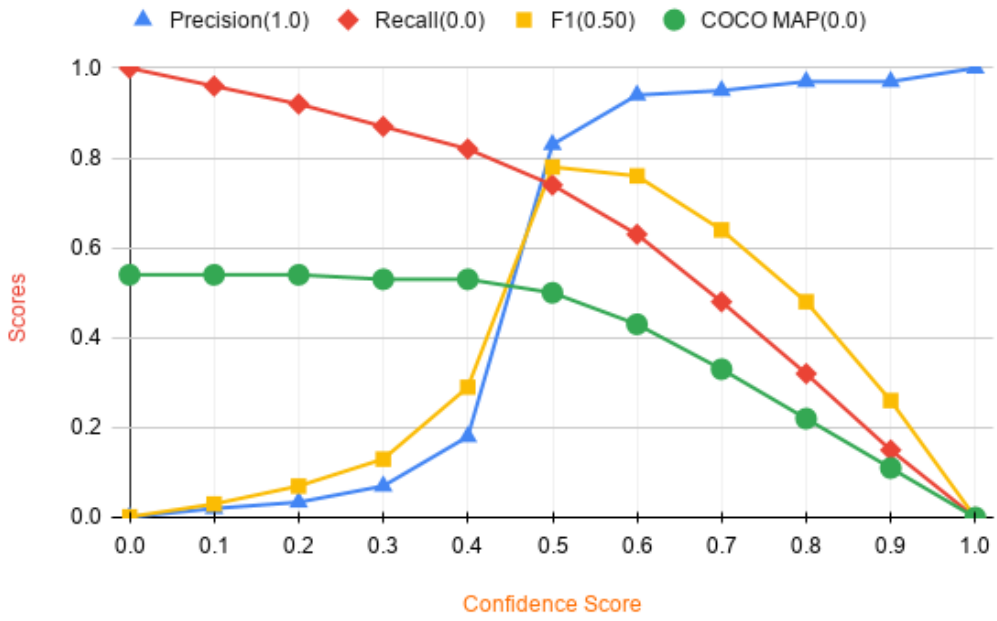

Figure 18: **Effect of thresholds on various metric at Val Set for 1-class CFSSD model**: Optimizing various metric for different thresholds. We find that for Precision, Recall and COCO MAP, optimum is obtained at trivial values of 1.0, 0.0 and 0.0 respectively, while for F1 score optimal value is obtained at 0.5.

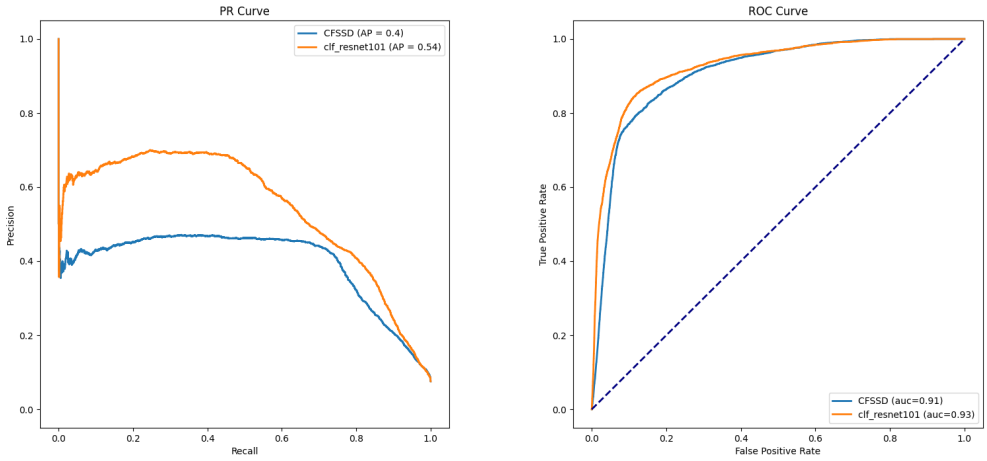

Figure 19: **PR and ROC curve for frame classification of the 1-class CFSSD Model on Test Set**

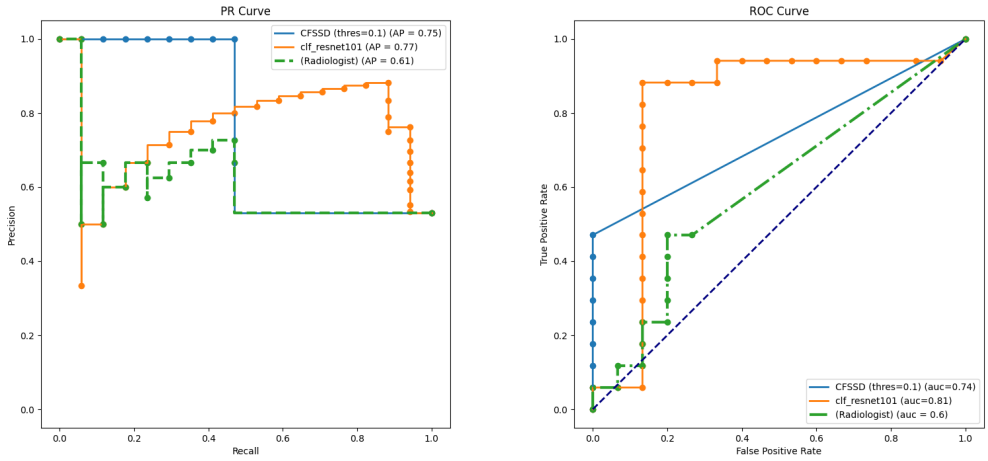

Figure 20: **PR and ROC curve for video level classification of the 1-class CFSSD Model on Test Set**

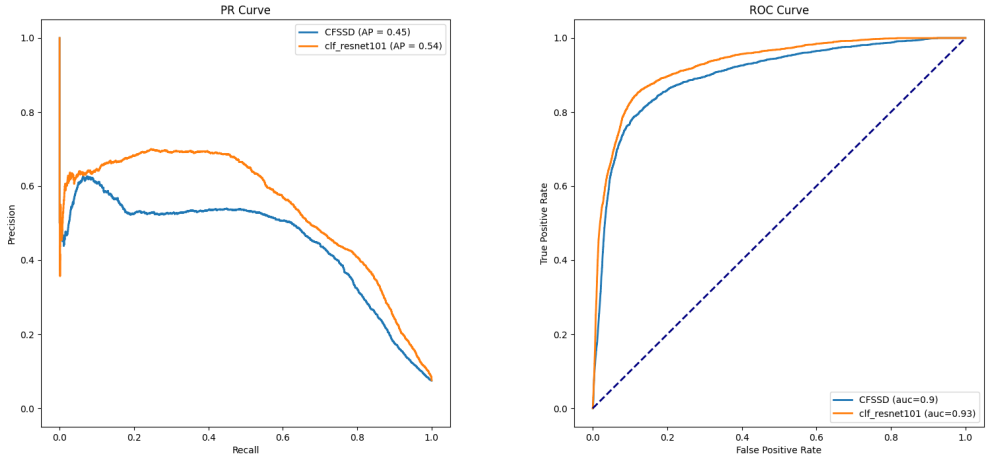

Figure 21: **PR and ROC curve for frame classification of the 1-class CFSSD Model on Test Set consisting of annotations from both annotators for both CFSSD and clf_resnet101**

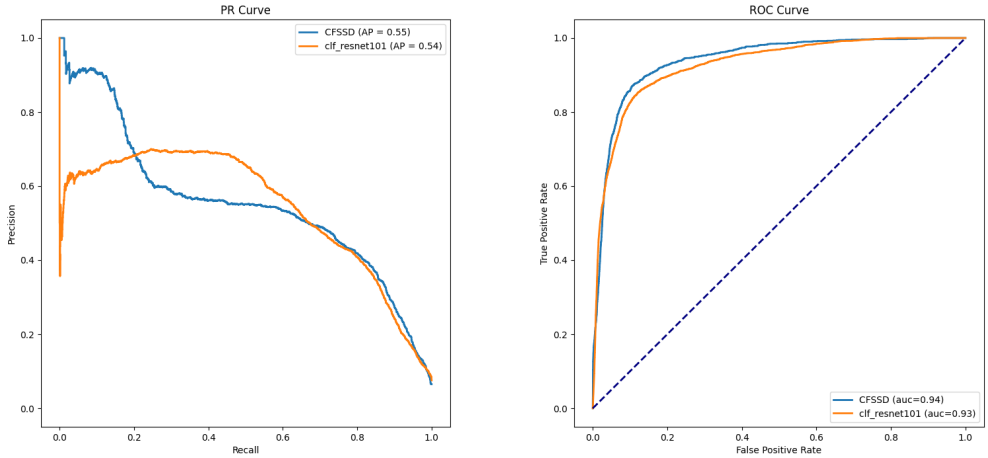

Figure 22: **PR and ROC curve for frame classification of the 1-class CFSSD Model on Test Set consisting of annotations from single annotators for CFSSD and both annotators combined for clf_resnet101**: It seems that CFSSD is doing better than the our baseline clf_resnet101, but important thing to note here is that our baseline contains the frames labels from both annotators but this model consists of the annotations from only one annotator for each frame.

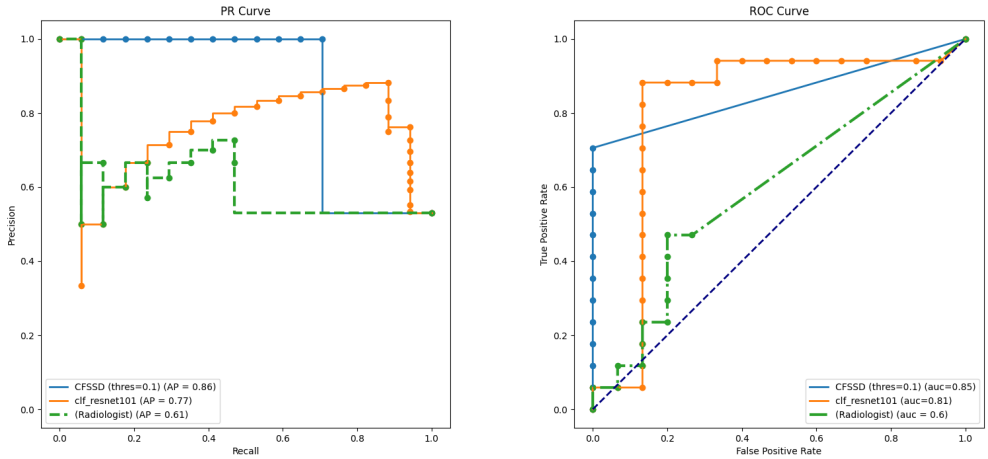

Figure 23: **PR and ROC curve for video level classification of the 1-class CFSSD Model on Test Set**

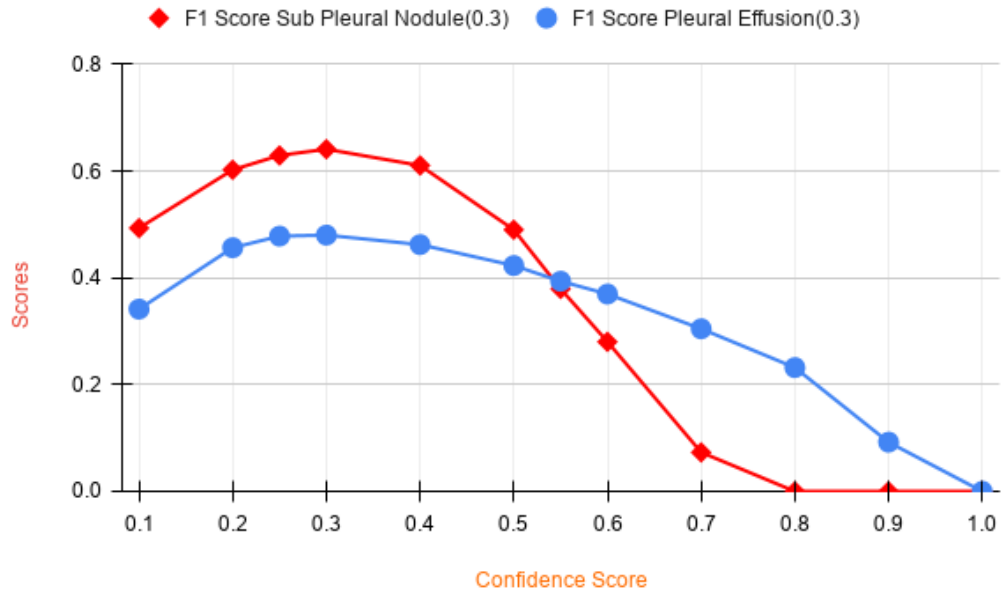

Figure 24: **Effect of thresholds on F1 Scores of Pleural Effusion and Sub Pleural Nodule at Val Set for 2-class SSD model**:Optimal values of the threshold is shown in the parenthesis in legend. On optimizing F1 score for different thresholds, we find that the optimal value is obtained at 0.3 for both Pleural Effusion and Sub Pleural Nodule.

Optimization of Confidence Score is done on the val set to get the maximum F1 Score for each of the two class. From figure (24) we see that optimal value for threshold on val set is 0.3. At this value we got the F1 score at val set to be **64.12% and 48.02%** for Sub Pleural Nodule and Pleural Effusion respectively. We used the same optimized threshold to calculate the F1 score on test set to get the F1 score of **32.93% and 55.48%** for Sub Pleural Nodule and Pleural Effusion respectively. Again the drop in performance is as expected since we have deliberately kept all the outliers in the test set.

5.7 2-class detection using SSD and only high quality annotations

Section 3.3 explains the High Quality datasets. These are the frames where boxes annotated by both annotators overlap with more than 50% IOU. In this experiment we used only the high quality frames.

We optimized the thresholds in the same way as done in the previous experiments on the F1 scores for Sub Pleural Nodule and Pleural Effusion on the val set. As seen in the figure (25), we found that the optimal value of threshold for Pleural Effusion is 0.26 and for Sub Pleural Nodule it is 0.56. We used this optimal thresholds to evaluate our model on Test Set and got the F1 score of 83.33% for Sub Pleural Nodule and 38.97% for Pleural Effusion.

5.8 1-class detection using SSD

In this case we again experimented with 1-class model but this time we did not include classification branch as in the case of CFSSD experiment. Here we only used the detection branch as done in the case of 2-class SSD explained in previous subsection. Our model outputs the classification score for one class only. After various experiments in this case VGG16[8] backbone turns out to be best performing model in contrast to all the previous cases where ResNet34[3] backbone gave the best result.

We optimized threshold to get the best F1 score for positive class and used the same threshold to evaluate the modelś performance on test set. We can see from figure (26) that the optimal value of confidence score(threshold) in this case is 0.36. The F1 score on val set at this threshold comes out to be 75.25% while at the test set F1 score for the same threshold comes out to be 66.12%.

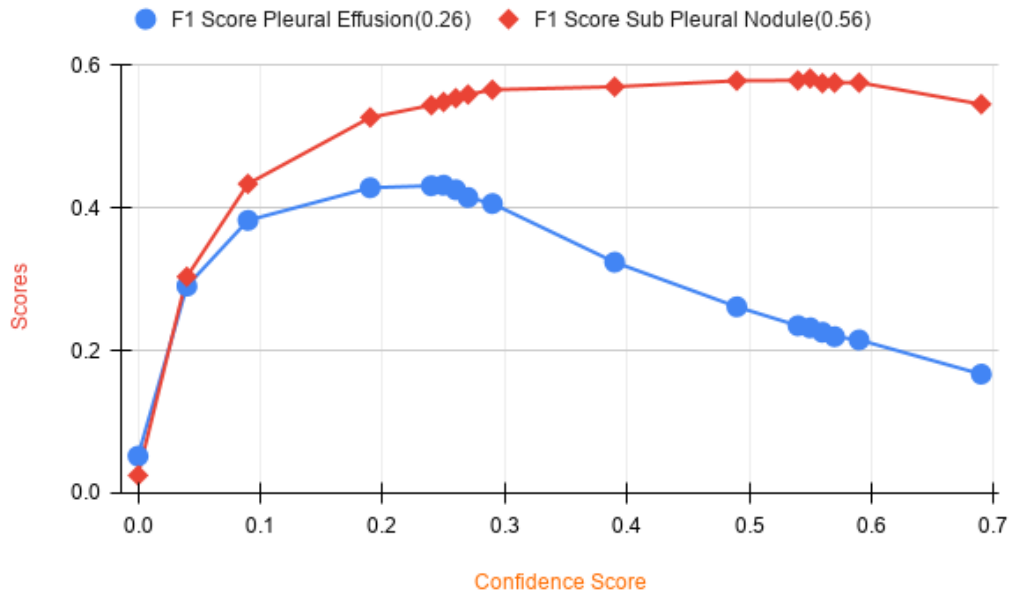

Figure 25: **Effect of thresholds on F1 Scores of Pleural Effusion and Sub Pleural Nodule at Val Set of High Quality dataset for 2-class SSD model**:Optimal values of the threshold is shown in the parenthesis in legend. On optimizing F1 score for different thresholds, we find that the optimal value is obtained at 0.56 for Sub Pleural Nodule and 0.26 for Pleural Effusion.

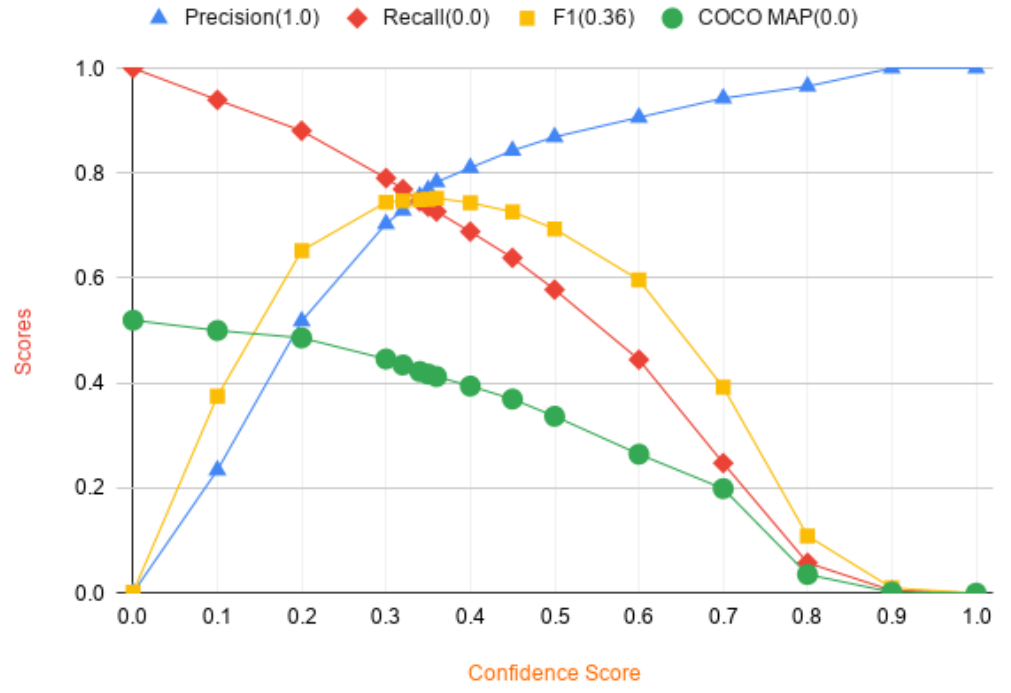

Figure 26: **Effect of thresholds on F1 Scores of positive class at Val Set of 1-class SSD model**:Optimal values of the threshold is shown in the parenthesis in legend. On optimizing F1 score for different thresholds, we find that the optimal value is obtained at 0.36.

6 Conclusion and Future Works

In this paper, we present a unique technique based on AI to solve one of the major problems related to the Tuberculosis, i.e. screening. Our method is not only safer than existing x-ray based screening techniques but it is also a Point-of-Care(POC) service with no involvement of any experts. Entire screening process can be carried by health care workers using our AI based tools. Through various experiments we showed that our AI based methods did better than radiologists and also provided explainability to its result.

The future work of this project includes resolving the problems in the datasets to whatever extent possible. Also, when we look at the video level data we have very less samples. Hence, increasing the amount of quality samples will be one of the major future task for this project. Since we have the videos of the patients, instead of modelling all frames independently, we plan to model this problem as an object tracking problem when we have more video samples.

References

[1] Katherine Floyd, Philippe Glaziou, Alimuddin Zumla, and Mario Raviglione. The global tuberculosis epidemic and progress in care, prevention, and research: an overview in year 3 of the end tb era. *The Lancet Respiratory Medicine*, 6(4):299–314, 2018.

[2] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[5] Zhi Zhen Qin, Melissa S Sander, Bishwa Rai, Collins N Titahong, Santat Sudrungrot, Sylvain N Laah, Lal Mani Adhikari, E Jane Carter, Lekha Puri, Andrew J Codlin, et al. Using artificial intelligence to read chest radiographs for tuberculosis detection: A multi-site evaluation of the diagnostic accuracy of three deep learning systems. *Scientific reports*, 9(1):1–10, 2019.

[6] 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, 2017.

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[8] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[9] Seok-Jae Heo, Yangwook Kim, Sehyun Yun, Sung-Shil Lim, Jihyun Kim, Chung-Mo Nam, Eun-Cheol Park, Inkyung Jung, and Jin-Ha Yoon. Deep learning algorithms with demographic information help to detect tuberculosis in chest radiographs in annual workers' health examination data. *International journal of environmental research and public health*, 16(2):250, 2019.

[10] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[11] Jianning Chi, Ekta Walia, Paul Babyn, Jimmy Wang, Gary Groot, and Mark Eramian. Thyroid nodule classification in ultrasound images by fine-tuning deep convolutional neural network. *Journal of digital imaging*, 30(4):477–486, 2017.

[12] Hailiang Li, Jian Weng, Yujian Shi, Wanrong Gu, Yijun Mao, Yonghua Wang, Weiwei Liu, and Jiajie Zhang. An improved deep learning approach for detection of thyroid papillary cancer in ultrasound images. *Scientific reports*, 8(1):1–12, 2018.

[13] Xiangchun Li, Sheng Zhang, Qiang Zhang, Xi Wei, Yi Pan, Jing Zhao, Xiaojie Xin, Chunxin Qin, Xiaoqing Wang, Jianxin Li, et al. Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. *The Lancet Oncology*, 20(2):193–201, 2019.

[14] Zhantao Cao, Lixin Duan, Guowu Yang, Ting Yue, and Qin Chen. An experimental study on breast lesion detection and classification from ultrasound images using deep learning architectures. *BMC medical imaging*, 19(1):51, 2019.

[15] Aman Dalmia, Jerome White, Ankit Chaurasia, Vishal Agarwal, Rajesh Jain, Dhruvin Vora, Balasaheb Dhame, Raghu Dharmaraju, and Rahul Panicker. Pest management in cotton farms: an ai-system case study from the global south. In *KDD(2020)*, pages 40–47, 2020.

[16] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[18] Forrest N Iandola, Song Han, Matthew W Moskewicz, Khalid Ashraf, William J Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and< 0.5 mb model size. *arXiv preprint arXiv:1602.07360*, 2016.

[19] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.