

©2020 Society of Photo-Optical Instrumentation Engineers (SPIE). One print or electronic copy may be made for personal use only. Systematic reproduction and distribution, duplication of any material in this paper for a fee or for commercial purposes, or modification of the content of the paper are prohibited. Access to this work was provided by the University of Maryland, Baltimore County (UMBC) ScholarWorks@UMBC digital repository on the Maryland Shared Open Access (MD-SOAR) platform.

Please provide feedback

Please support the ScholarWorks@UMBC repository by emailing scholarworks-group@umbc.edu and telling us what having access to this work means to you and why it's important to you. Thank you.

PROCEEDINGS OF SPIE

[SPIDigitalLibrary.org/conference-proceedings-of-spie](https://spiedigitallibrary.org/conference-proceedings-of-spie)

Active semi-supervised expectation maximization learning for lung cancer detection from Computerized Tomography (CT) images with minimally label training data

Nguyen, Phuong, Chapman, David, Menon, Sumeet, Morris, Michael, Yesha, Yelena

Phuong Nguyen, David Chapman, Sumeet Menon, Michael Morris, Yelena Yesha, "Active semi-supervised expectation maximization learning for lung cancer detection from Computerized Tomography (CT) images with minimally label training data," Proc. SPIE 11314, Medical Imaging 2020: Computer-Aided Diagnosis, 113142E (16 March 2020); doi: 10.1117/12.2549655

SPIE.

Event: SPIE Medical Imaging, 2020, Houston, Texas, United States

Active Semi-Supervised Expectation Maximization Learning for Lung cancer Detection from Computerized Tomography (CT) images with Minimally Labeled Training Data

Phuong Nguyen^{*a}, David Chapman^{**a}, Sumeet Menon^a, Michael Morris^b, Yelena Yesha^a

^aDept. of Computer Science and Electrical Engineering, University Of Maryland Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250; ^bDiagnostic Radiology, Nuclear Medicine, and Internal Medicine, Mercy Medical Center, 345 St. Paul Place, Baltimore, MD 21202

ABSTRACT

Artificial intelligence (AI) has great potential in medical imaging to augment the clinician as a virtual radiology assistant (vRA) through enriching information and providing clinical decision support. Deep learning is a type of AI that has shown promise in performance for Computer Aided Diagnosis (CAD) tasks. A current barrier to implementing deep learning for clinical CAD tasks in radiology is that it requires a training set to be representative and as large as possible in order to generalize appropriately and achieve high accuracy predictions. There is a lack of available, reliable, discretized and annotated labels for computer vision research in radiology despite the abundance of diagnostic imaging examinations performed in routine clinical practice. Furthermore, the process to create reliable labels is tedious, time consuming and requires expertise in clinical radiology. We present an Active Semi-supervised Expectation Maximization (ASEM) learning model for training a Convolutional Neural Network (CNN) for lung cancer screening using Computed Tomography (CT) imaging examinations. Our learning model is novel since it combines Semi-supervised learning via the Expectation-Maximization (EM) algorithm with Active learning via Bayesian experimental design for use with 3D CNNs for lung cancer screening. ASEM simultaneously infers image labels as a latent variable, while predicting which images, if additionally labeled, are likely to improve classification accuracy. The performance of this model has been evaluated using three publicly available chest CT datasets: Kaggle2017, NLST, and LIDC-IDRI. Our experiments showed that ASEM-CAD can identify suspicious lung nodules and detect lung cancer cases with an accuracy of 92% (Kaggle17), 93% (NLST), and 73% (LIDC) and Area Under Curve (AUC) of 0.94 (Kaggle), 0.88 (NLST), and 0.81 (LIDC). These performance numbers are comparable to fully supervised training, but use only slightly more than 50% of the training data labels.

Keywords: Lung Cancer screening, Active Learning, Semi-Supervised Learning, CT, Label Acquiring, Computer-Aided Diagnosis, Expectation Maximization, Artificial Intelligence, Deep Learning

1. INTRODUCTION

Deep learning using Convolutional Neural Networks (CNNs) has greatly improved the performance of Computer Aided Diagnosis (CAD) algorithms for cancer screening in recent years [1, 2, 3, 4, 5, 6, 7]. However, a disadvantage of many deep learning classification techniques including many CNNs is that these algorithms are fully supervised and therefore require very large datasets with manual annotation by expert radiologists in order to achieve high accuracy. Typically these fully annotated datasets are on the order of thousands of images, whereas clinical Picture Archiving and Communication Systems (PACS) even at a community hospital contains millions of unlabeled or weakly labeled Radiology examinations. As such, a major challenge in applying deep learning based CAD clinically is to be able to make use of larger *unlabeled* radiology imaging datasets and to combine these datasets with smaller highly annotated datasets.

* phuong3@umbc.edu 410-455-3535 ; ** dchapm2@umbc.edu

Methods to reduce the amount of manual annotation necessary while maintaining or improving accuracy are an important contribution because manual annotation for medical imagery is time consuming, costly, and requires expert labelers with a high level of expertise in radiology.

Accurate Chest CT annotation for lung cancer screening requires a board certified diagnostic radiologist (4 years of medical school, 1 year of internship, and 4 years of diagnostic radiology residency), ideally with additional experience or subspecialization in Thoracic Radiology or Oncologic Radiology (1-2 years additional fellowship or clinical experience). Furthermore, challenging tasks such as nodule segmentation and malignancy assessment require additional annotation that is beyond the routine clinical standard of care and therefore is not readily available. The misclassification rate for CNNs has been empirically estimated to decay exponentially as data volumes increase [8]. Therefore, the absence of large annotated datasets are currently a limiting factor in the clinical application of deep learning for radiology. Active learning and Semi-supervised learning algorithms have the potential to enable deep learning based CAD to further improve performance by making use of large clinical image volumes collected by institutions thereby greatly reducing the necessary labeling burden.

In this study we investigate a novel learning model that combines Active learning with Semi-supervised learning in order to reduce the amount of annotation necessary to create a CNN based CAD algorithm for Chest CT cancer screening examinations. Lung cancer screening was recently identified as contributing to the largest year-over-year decline in cancer deaths ever recorded. [9] We demonstrate that an Active Semi-Supervised Expectation Maximization (ASEM) algorithm is a viable approach for training deep CNN based CAD algorithms using CT exams. This work builds upon our recent work demonstrating that a Semi-supervised EM (SEM) algorithm was able to improve cross-validated Lung Cancer screening accuracy as compared to a fully-supervised technique [10]. We expand on this algorithm by incorporating Active learning in combination with Expectation-Maximization (EM) in order to further improve cross-validated screening accuracy. The active learning component allows the algorithm to interactively suggest images to radiologists that need to be labeled, and the semi-supervised learning using EM allows the algorithm to incorporate a larger unlabeled training image dataset along with a smaller labeled dataset. The suggested images from the the large unlabeled pool are selected by a validating classification uncertainty method.

2. RELATED WORK

Recently, [7] presented an artificial intelligence (AI) system which can potentially outperform human experts in breast cancer prediction. To evaluate its effectiveness in the clinical setting, they have curated a large characteristic dataset from the UK and large enhanced dataset from the USA and have manifested complete reduction of 5.7% (USA) and 1.2% (UK) in false positive rate and 9.4% (USA) and 2.7% (UK) in false negative rate. Expectation Maximization (EM) is an influential generative meta-algorithm for latent variable training and has been employed for semi-supervised learning [11]. Generative algorithms model the probability distribution of unlabeled imagery as a function of model and labeled imagery [12 , 13]. Although EM assumes an underlying generative model, recent work in combining EM with discriminative CNN architectures have been shown to be successful in practice, likely due to the non-linearity of CNNs. EM is applied to improve semantic segmentation of general imagery using CNNs [14]. The method achieves 73.9% accuracy with a small number of pixel-level annotated images which is almost competitive with the fully-supervised model's accuracy of 79%.

Active learning or "optimal experimental design" in statistics is part of the machine learning field, where the learner selectively asks (or queries) experts for more ground truth labels in order to achieve its desirable outcome (e.g model's accuracy or better learning with less samples). As such, Active learning methods choose the most informative unlabeled samples for annotation by a human radiologist. The selection process requires the learning algorithm to provide a query strategy to select unlabeled data points that are most likely to improve the model accuracy if labeled. By using uncertainty sampling [15], the way to select is by picking the least certain label and requiring experts to annotate. Recently, active learning has been to overcome data scarcity issues with current models by incrementally choosing the most revealing unlabeled samples, querying their labels and putting them to the labeled data set [16].

The Monte Carlo dropout method is used to estimate the level of uncertainty in the active learning process or look ahead technique to select samples [17]. Better uncertainty estimation is obtained using ensemble models [18]. Previous work combined the approach proposed by [17] and data augmentation (generate new training samples from a latent variable, discriminate between real and fake samples) for classification learning tasks [19].

Active learning alone has been applied for characterization of endothelial cells in human tumors [20] and predicting positive p53 cancer rescue regions by using the most informative information method [21]. There is a recent active learning framework that is presented by [22] for skin lesion analysis which is cost-effective by selecting and employing much fewer labeled samples while the network still attains state-of-the-art performance. Their active learning method tends to enhance the annotation coherence. The authors have selected their samples to be highly supportive and have used dataset of the ISIC 2017 Skin lesion Classification challenge and attained state-of-the-art performance by using 50% of the data for the first task and 40% of the data for the second task of skin lesion classification. Previous work in [23] developed a model that recognizes anomalies within plain-text-based reports which could then be utilized further as a method to create labels for models depending on CT scans thereby aiming to decrease human efforts in labeling CT scans. A systematic approach named NoduleX was proposed by [24] which uses a deep learning CNN as well as a radiomics approach for prediction of lung nodule malignancy using CT images of the LIDC dataset.

3. METHODOLOGY

The ASEM method combines both Semi-supervised and Active learning to improve the accuracy of the model prediction with as few known labels as possible. Active learning techniques require an Oracle step in which the algorithm asks for more ground truth from the unlabeled data during the ASEM process. Figure 1 shows an overview of our proposed learning model. We first train an initial model using a subset of the training data which is fully labeled. Subsequently, the ASEM model alternates between Semi-supervised *Expectation* and *Maximization* steps as well as Active learning *Oracle*, and *Active Retraining* steps. Each ASEM iteration requires retraining the model in the *Maximization* and *Active Retraining* steps with improved estimates of the latent variables either due to *Expectation*, or due to the *Oracle*. The computational burden of retraining the model for each ASEM iteration however is greatly reduced by re-using the weights from the previous ASEM iteration rather than retraining from a random seed (see table 4 for the runtime performance of our ASEM-CAD).

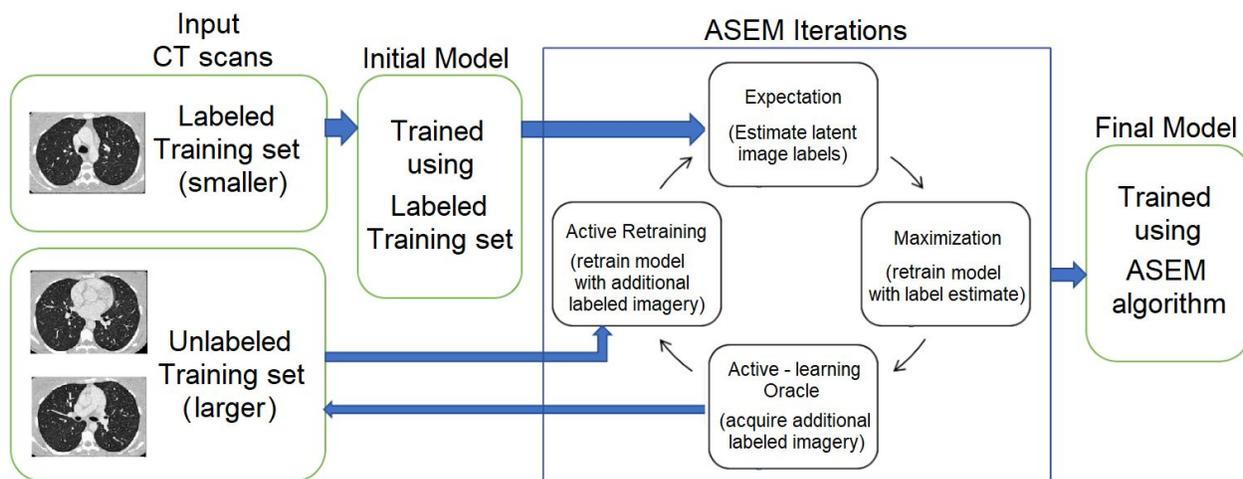


Figure 1 Overview of proposed model for lung cancer detection

The ASEM algorithm is an alternating local maximization-maximization algorithm in which we attempt to perform a maximum likelihood estimate of our model θ and expected experimental design ξ in the presence of latent variables Z and given the ability to actively label a finite number of observations y . Our goal is to show that the maximum likelihood of the model is improving after each Active Learning step and after each EM step.

A theoretical detail is that EM steps attempt to maximize likelihood whereas Active learning minimizes cross entropy. These have equivalent global optima under the assumption of statistical independence and have *approximately equivalent* optima in practical machine learning applications as follows,

$$-\sum_i p(X_i) \log(p(X_i | \theta)) \approx -\log(p(X|\theta)) \quad (1)$$

The *Expectation* and *Maximization* steps maximize likelihood of the model under all possible values of the latent variable space [11]. The likelihood of a latent variable is given by the integral of the joint probability density over all possible values of the latent variable Z .

$$L(\theta; X) = p(X|\theta) = \int p(X, Z|\theta) dZ \quad (2)$$

EM attempts to solve the above integral by alternating between Expectation and Maximization steps. Expectation is in which we calculate the expected value of the latent variables given the t^{th} iteration of the model θ^t . In the context of a deep learning framework, the expected value of $E_{Z|X, \theta^t}$ can be computed by classifying label probabilities of the unlabeled imagery using the t^{th} iteration of the model coefficients θ^t .

$$Q(\theta|\theta^t) = E_{Z|X, \theta^t} [\log L(\theta; X, Z)] \quad (3)$$

The Maximization step is to compute the maximum likelihood model θ^{t+1} given our current expected value of the latent variables Z . This can be accomplished by retraining the deep learning model using the expected value of the image labels at the t^{th} iteration.

$$\theta^{t+1} = \operatorname{argmax}_{\theta} Q(\theta|\theta^t) \quad (4)$$

Our active learning process is designed to select data points for an expert human to label during the EM iteration training. In this way we make selective incremental improvements to data label quality. The active learning steps optimize the expected posterior cross entropy of the model given an alternate experimental design ξ with the addition of a labeled sample y_i . We cannot measure the posterior cross entropy directly because we must select a sample before acquiring its true label via Oracle. As such the expected posterior cross entropy is as follows

$$U(\xi) = -\int \log(p(X|\theta, y_i, \xi)) dy_i \quad (5)$$

This quantity can be rewritten using Bayes rule for bayesian experimental design as follows,

$$U(\xi) = -\int \log(p(y_i|\theta, X, \xi) \frac{p(y_i|\theta, \xi)}{p(X|\theta, \xi)}) dy_i \quad (6)$$

This integral would be expensive to compute as it would require retraining the algorithm for every possible sample choice and every possible sample label prior to choosing the appropriate sample.

However, we can make an approximation that a single sample does not change the model prediction of *most samples* more than a small amount at a time, but rather the predicted sample itself y_i has the greatest local contribution to posterior cross entropy.

Under this assumption the change in posterior cross entropy is approximately equal to the normalized classification entropy over all possible K labels as follows,

$$\Delta U(\xi) \approx I_{norm}(y_i) = \frac{-1}{\log(K)} \sum_{k=1}^K p(y_{ik}) \log(p(y_{ik})) \quad (7)$$

We can perform a small number of Active Learning steps in a batch rather than performing a single active learning step. In this case, our expected utility becomes the normalized classification entropy of all of the selected samples in the batch as follows,

$$avg(I_{norm}(Y)) = \frac{1}{|Y|} \sum_{y \in Y} I_{norm}(y) \quad (8)$$

As such the ASEM algorithm alternates between steps 3, 4, and 8 in order to improve the maximum likelihood estimate and reduce classification cross entropy in the presence of latent variables while optimizing Bayesian experimental design. An important point is that algorithm, as a maximization-maximization meta-algorithm does not guarantee convergence to a global optimum, but rather will achieve a local optimal experimental design as well as a local optimal estimate of latent variables for semi-supervised learning.

4. DATA AND EXPERIMENTAL DESIGN

We analyze the performance of the active semi-supervised EM (ASEM) model using 3 lung cancer screening datasets: Kaggle, NLST, and LIDC-IDRI. The Kaggle Data Science bowl (2017) (Kaggle17) is a benchmark dataset for Computer Aided Diagnosis (CAD) algorithms for Non-Small Cell Lung Cancer (NSCLC) cancer screening using Low Dose Computed Tomography (LDCT) scans. Each volumetric scan contains varying number of Chest CT image slices, and each slice is the standard resolution of 512x512 pixels. We experiment with the Kaggle17 dataset, which consists of a total of 1375 patients. These scans are labeled as 1 for cancerous (i.e. diagnosed with lung cancer within one year of the scan) and 0 for non-cancerous. NLST was a landmark 2011 study that proved that high risk individuals (60+ yrs old, and heavy smokers) who receive periodic LDCT lung cancer screening exams have greater life expectancy and lower mortality than if these individuals were to receive periodic chest x-ray screenings. We used 4075 LDCT scans from the NLST dataset, and each scan was labeled as 1 if the patient was diagnosed with cancer or 0 if the patient was not diagnosed with cancer. Of these 4075 scans, 639 patients were diagnosed with lung cancer.

The LIDC-IDRI Dataset [25, 26] is a publicly available dataset that consists of diagnostic and lung cancer screening thoracic computed tomography (CT) scans with marked-up annotated lesions. This dataset is a web-accessible international resource initiated by the National Cancer Institute (NCI), and then further developed by the Foundation for the National Institutes of Health (FNIH) and going along with the Food and Drug Administration (FDA). LIDC is used for the purpose of research towards development, training and assessing of computer-aided diagnostic (CAD) methods for detecting and diagnosing lung cancer in its early stages. This dataset is created in collaboration with seven academic centers and eight medical imaging companies that have 1018 CT cases where it has thoracic CT scans associated with an XML file. The LIDC study has annotations that are provided by four experienced

Thoracic Radiologists who reviewed each of the 1018 CT cases in the LIDC/IDRI cohort and marked lesions into 3 categories based on the nodule size. Nodules ≥ 3 mm have a greater probability of being malignant than nodules < 3 mm and non-nodule ≥ 3 mm. The malignancy rating is given from 1-5 depending on the size and features of the nodule.

The following section presents how we build and evaluate our ASEM model for Computer Aided Diagnosis, called ASEM-CAD for lung cancer detection using the above 3 datasets.

4.1 Data Pre-processing

In Both Kaggle and NLST datasets, each patient CT scans have varied slice numbers. For each patient, we create standard 3D volume data as input for the model by resizing the 512x512 image pixels of multiple DICOM slices into a standard 50x50x20 voxel resolution. The third dimension is reduced to 20 by chunking slices into 20 chunks then average. Thus, the input 3D volume for each patient has 50x50x20 dimensions, each associated with a label: either 1 (cancer) or 0 (non-cancer). Kaggle has 1357 patients with 356 cancer cases. Cancer cases represent 26% of the Kaggle dataset and non-cancer cases represent 74%. For NLST, after preprocessing we have 2538 cases, 397 cancers and 2141 non-cancer.

The third dataset, LIDC, has 1010 CT scans, with each slice having 512x512 pixels. We crop 4253 nodules which cover the size of nodules according to annotations that are provided by the four experienced radiologists. Their annotations are given in the form of nodule Region Of Interest and their Z-positions. Thus we crop the the 32x32x16 dimension nodules using the spatial coordinate centered at the annotated location of the CT scans. For assigning the labels of the nodule, we use rating scores provided by board certified radiologists with levels 1 and 2 as a non-cancer nodule (benign), level 4 and 5 as cancer nodule (malignant). The score level 1 meaning highly benign, 2 as moderately benign (non-cancer), 4 as moderately suspicious to be malignant and 5 as highly likely to be malignant (cancer). Nodules labeled by a radiologist as having an intermediate malignancy (rating 3) are not considered for classification in this paper. In summary, we have 4253 nodules, each has 32x32x16 dimensions and an associated label 1 (cancer) and 0 (non-cancer). There are 1653 cancerous nodules. 2600 nodules belong to benign cases.

4.2 Neural Architecture and Training Procedures

The ASEM-CAD neural architecture has six CNN layer blocks. Each has Convolutional 3D layers, LeakyRelu, BatchNormalization, MaxPooling3D, DropOut with 32, 32, 64, 64, 64, 64 feature maps. The Convolutional 3D uses a 3x3x3 filter. Then follow by Dense, BatchNormalization, DropOut layer with 256 features. The last layer has Dense 1024 and 2 classes. For the LIDC dataset, we use a simpler CNN architecture with layer block 1, 2, 3, 4, 7 and 8. The CNN feature maps are 8, 8, 16, 16 for LIDC experiments. For all experiments, data is splitting as 80% of input dataset is used for ASEM to train and evaluate the model. The remaining 20% of dataset is used for testing.

The ASEM training procedure is as follows: the initial model is fully trained with 50% of all labels until convergence using category Cross-entropy loss. The ASEM-CAD is trained using RMSprop optimizer with a learningRate of 0.0001. The Initial model is fully trained using 500 epochs. Then each ASEM meta-iteration (EM iteration) is trained with 10 epochs. The ASEM model is trained in batches of 32 samples. The Initial model is saved.

Then EM iterations can load the initial model's parameters and start the Active EM training. The Active component selects 10 samples and asks an Oracle for the label during an ASEM iteration. The number of ASEM' Active EM iterations is set to 5. In addition, we apply Label Smoothing, BatchNormalization, and Early Termination techniques for training our ASEM-CAD.

5. RESULTS

In this section, we present our experimental findings of the performance of the ASEM algorithm in comparison to fully supervised learning as well as in comparison to the SEM algorithm. We calculate Receiver Operating Curves (ROC), and present accuracy, Area Under Curve (AUC), sensitivity, specificity, and precision as evaluation metrics. We evaluate the ASEM-CAD using the Kaggle17, NLST, and LIDC datasets and compare with fully supervised training as well as Semi-Supervised EM. We compare the following methods,

Supervised 1: Using only 50% of these labeled datasets;

Supervised 2: Using 100% labeled dataset;

SEM-CAD: Start with 50% labeled data initially then using full dataset for EM iterations.

ASEM-CAD 1: Active semi-supervised with 50% labels and additional labels with Max. Classification Entropy;

ASEM-CAD 2: Active semi-supervised with 50% labels, add additional labels with above Avg Classification Entropy.

Most lung cancer datasets including Kaggle17, NLST, and LIDC have an unbalanced number of cancer vs non cancer cases with a greater number of non-cancerous cases relative to cancerous cases. Yet in clinical practice, it is necessary to bias the final threshold of any cancer screening test such as to over-predict false positives in order to reduce the probability of predicting false negatives. In order to provide a more complete picture, ROCs are calculated by varying the prediction threshold between 0 to 1 and plotting *sensitivity* against $1 - specificity$ if all predictions above the varied threshold are classified as cancerous. AUC varies from 0 to 1 (higher is better), and is defined as the integral of *sensitivity* with respect to $1 - specificity$ over the domain of the ROC curve. Tables 1, 2, and 3 calculate an inflection point along this ROC curve and present *sensitivity*, *specificity*, and *precision*. Table 1 shows the performance of the ASEM-CAD algorithm over the Kaggle17 dataset, Table 2 shows the performance over the NLST dataset, and Table 3 shows the performance over the LIDC-IDRI dataset.

Table 1. ASEM performance over the Kaggle17 dataset

Experiments	Number of Samples	Test_ACC	AUC	Sensitivity	Specificity	Precision
Supervised 1	50% labels only	0.87	0.85	0.69	0.94	0.79
Supervised 2	100% labels	0.92	0.92	0.81	0.95	0.85
SEM-CAD	50% label initially	0.91	0.92	0.88	0.92	0.76
ASEM-CAD1	50%, add labels with Max. Classification Entropy	0.92	0.94	0.78	0.96	0.88
ASEM-CAD2	50%, add labels with above Avg Classification Entropy	0.85	0.81	0.66	0.9	0.67

Over the Kaggle17 dataset, ASEM-CAD1 outperformed Supervised 2 algorithms in AUC (0.94 vs 0.92), this is notable because Supervised 2 has the benefit of using 100% of the training labels.

At the inflection point, Specificity and Precision were higher although Sensitivity was slightly lower as compared to fully-supervised learning model using 100% labels (table 1). It also showed in table 1 that ASEM-CAD1 outperforms the SEM-CAD in similar metrics. Noticeably, ASEM-CAD1 performed much better by 7.9%, 13%, 15%, 6%, and 23% in all metrics respectively in its order in table 1 over our ASEM-CAD2. ROC curves comparing Supervised 2, SEM-CAD, and ASEM-CAD1 are shown in Figure 2.

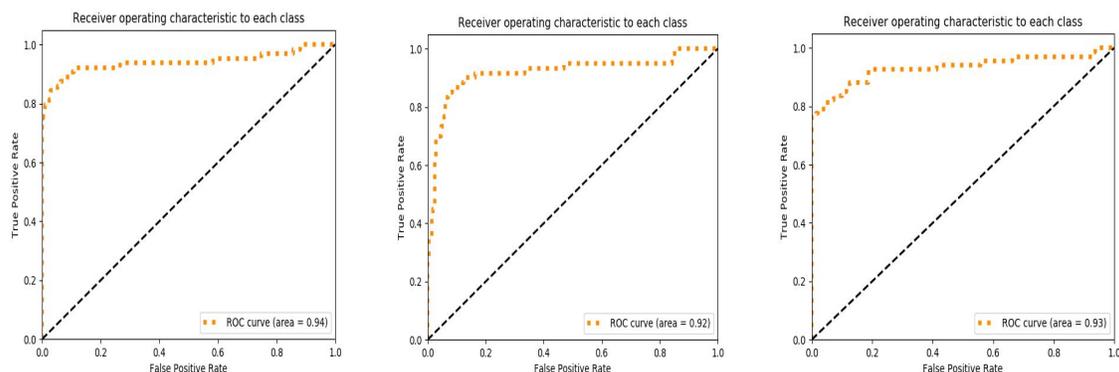


Figure 2 ROC analysis of the Kaggle dataset **a) left Supervised 2** (100% labels) **b) middle semi-supervised SEM-CAD** (50% labels). **c) right our ASEM-CAD1**, Active Semi-Supervised (50% labels, add labels with Max. Classification Entropy). Note: this ROC curve is reported per run not on average of multiple runs presented in table 1.

Table 2 shows a similar ROC analysis of ASEM-CAD1 against supervised and semi-supervised techniques. We see that the ASEM-CAD1 algorithm achieves AUC of 0.88 which is comparable and slightly greater than Supervised 2 which achieves AUC of 0.87. Supervised 2 has the benefit of using all of the labeled data, whereas ASEM-CAD1 uses only using slightly more than half of the labeled data. Figure 3 shows the ROC curves comparing ASEM-CAD2 versus Supervised 2, we see that ASEM-CAD2 exhibits comparable performance characteristics to Supervised 2 in addition to achieving similar AUC.

Table 2. ASEM performance over the NLST dataset.

Experiments	Number of Samples	Test_AC C	AUC	Sensitivity	Specificity	Precision
Supervised 1	50% labels only	0.92	0.87	0.65	0.97	0.75
Supervised 2	100% labels	0.94	0.87	0.72	0.97	0.90
SEM-CAD	50% label initially	0.90	0.89	0.77	0.92	0.67
ASEM-CAD1	50%, add labels with Max. classification entropy	0.93	0.88	0.56	0.99	0.94
ASEM-CAD2	50%, add labels with above Avg classification entropy	0.92	0.86	0.63	0.99	0.91

Table 3 compares the performance of ASEM versus supervised and semi-supervised methods for nodule malignancy estimation using the LIDC-IDRI dataset.

We see that ASEM-CAD1 achieves AUC of 0.81 which is very comparable performance to Supervised 2 (AUC 0.82), by using only slightly more than 50% of the data labels, as opposed to 100% of the data labels. For this dataset ASEM-CAD1 and ASEM-CAD2 achieved comparable AUC performance and these algorithms outperformed SEM-CAD. At the inflection point ASEM-CAD1 achieves slightly greater sensitivity but slightly lower specificity than Supervised 2. We compare the ROC curves for Supervised 2 vs ASEM-CAD1 in Figure 4, and we find that these curves have similar accuracy performance characteristics.

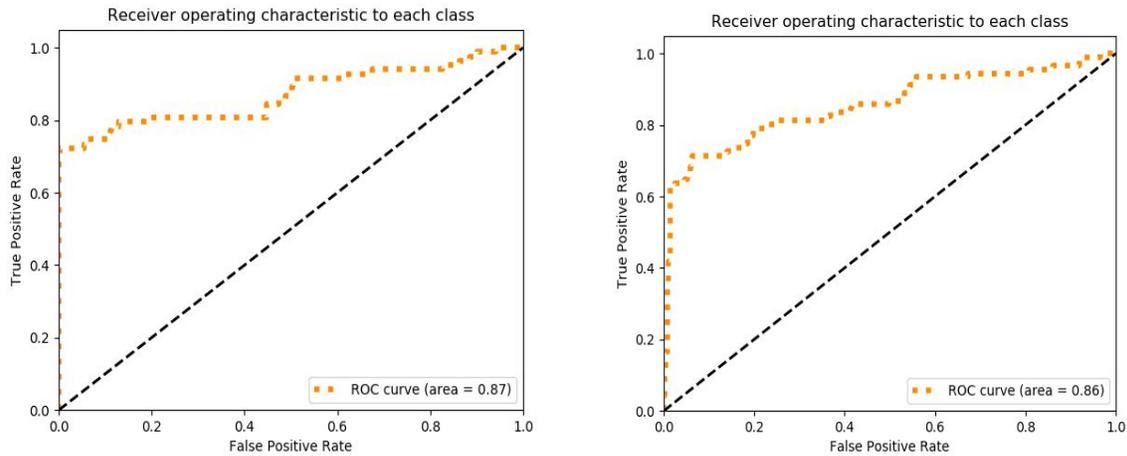


Figure 3 ROC analysis of the NLST dataset a) left fully-supervised(100% labels). b) Right ASEM-CAD2 Active Semi-supervised (50% labels), add labels with above Average Classification Entropy.

Table 3. ASEM performance over the LIDC-IDRI dataset.

Experiments	Number of Samples	Test_ACC	AUC	Sensitivity	Specificity	Precision
Supervised 1	50% labels only	0.67	0.82	0.91	0.52	0.54
Supervised 2	100% labels	0.74	0.82	0.78	0.72	0.64
SEM-CAD	50% label initially	0.71	0.81	0.82	0.64	0.59
ASEM-CAD1	50%, add labels with Max. classification entropy	0.73	0.81	0.79	0.70	0.62
ASEM-CAD2	50%, add labels with above avg classification entropy	0.73	0.80	0.82	0.67	0.60

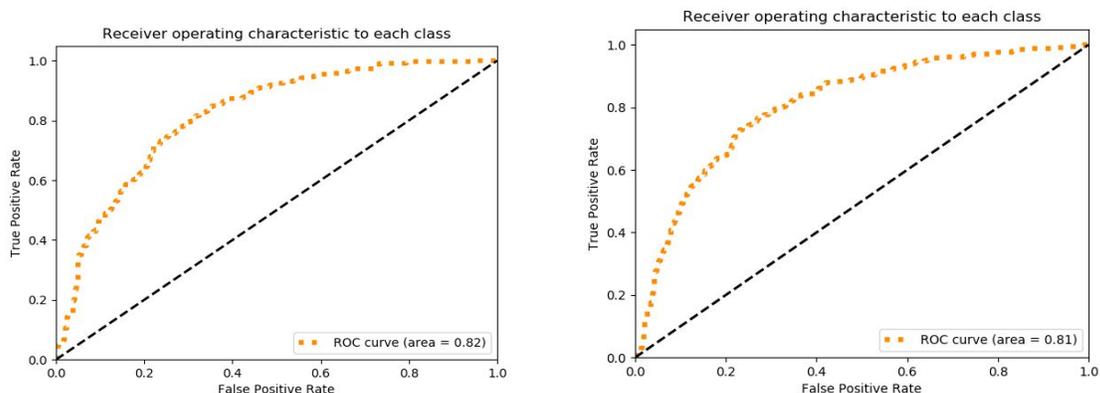


Figure 4 ROC analysis of the LIDC dataset a) left fully-supervised (100% labels)-Supervised2. b) Right Active Semi-supervised (50% labels, add label with Maximization Classification Entropy) ASEM-CAD1.

Also notable is that the ASEM algorithm, despite iteratively retraining of the CNN models more than 10 times, adds less than 50% additional overhead to the overall training time. Table 4 shows the wall-time runtimes of the ASEM-CAD algorithm for training using a customized built computer with AMD 1885 MHz 32 cores, 658 GB, 3 NVIDIA GeForce RTX, each GPU has 11 GB memory. The reason that the runtime is manageable (30%-50% increase) as opposed to a factor 10x or more is because we save and reuse the CNN weights after each iteration as opposed to retraining the CNN from random weights. As a maximization-maximization procedure, each ASEM iteration can be thought of as a local hillclimb in order to further improve the maximum likelihood estimate of the model parameters. Thus, the weights from the previous ASEM step a good initial guess to the weights of the subsequent ASEM step, thereby reducing the number of epochs necessary (and thus total walltime) for the ASEM iterations.

Table 4 Training wall time of the ASEM-CAD algorithm

Dataset	Number of Images	Total runtime in Minutes			
		Initial Model	ASEM Iterations	Total Time	Percent Increase
Kaggle	1357	26	8	34	31 %
NLST	2538	47	15	62	32 %
LIDC	4253	8	4	12	50 %

5. CONCLUSIONS

ASEM-CAD is a new CNN based CAD model which combines both semi-supervised and active learning to detect lung cancerous nodules and lung cancer cases using CT scans, while reducing the number of labeled scans necessary to train the neural architecture. ASEM-CAD has been evaluated using three public chest CT datasets for lung-cancer screening: Kaggle17, NLST, LIDC. Our experiments showed ASEM-CAD can detect lung cancer with high AUC performance comparable to that of fully supervised learning, but with only slightly more than 50% of the training labels. The ASEM-CAD1 vs Supervised2 AUC performances were: NLST (0.88 vs 0.87), Kaggle17 (0.94 vs 0.92), and LIDC-IDRI: (0.81 vs 0.82). The Active learning component asks for additional ground truth of unlabeled data which has a high level of classification uncertainty (high entropy) during the EM training process. This selection process results in better performance as compared to purely Semi-Supervised learning as well (SEM-CAD).

In conclusion, we have demonstrated that ASEM-CAD is able to detect suspicious lung nodules with comparable accuracy as using a fully supervised algorithm but with far fewer labeled images. ASEM-CAD may help to provide medical imaging researchers and commercial vendors with a more practical approach to train more powerful artificial intelligence based virtual radiology assistants (vRA) to augment radiologists interpreting oncologic imaging in the setting of lung cancer screening and perhaps other diagnostic radiology examinations more generally. In the future, we expect that Semi-Supervised and Active learning will play an increasingly larger role in the development of Deep CAD algorithms as these techniques will make it possible to learn from large clinical PACS datasets while reducing the need for manual annotation by radiologists.

ACKNOWLEDGEMENTS

This work was sponsored by NSF IUCRC University of Maryland Baltimore County: Center for Accelerated Real Time Analytics (CARTA), <https://carta.umbc.edu/>. We would like to thank Jayalakshmi Mangalagiri for helping in preparing this manuscript. Additional thanks to Kushal Mehta, Arshita Jain, and Ankita Viresh Rathod for pre-processing the LIDC dataset. Special thanks to Eliot Siegel for his contributions to our research related to CAD algorithms for cancer screening.

REFERENCES

- [1] Litjens, Geert, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I. Sánchez. "A survey on deep learning in medical image analysis." *Medical image analysis* 42 (2017): 60-88.
- [2] Bejnordi, Babak Ehteshami, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak et al. "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer." *Jama* 318, no. 22 (2017): 2199-2210.
- [3] Wang, Dayong, Aditya Khosla, Rishab Gargeya, Humayun Irshad, and Andrew H. Beck. "Deep learning for identifying metastatic breast cancer." *arXiv preprint arXiv:1606.05718* (2016).
- [4] Hua, Kai-Lung, Che-Hao Hsu, Shintami Chusnul Hidayati, Wen-Huang Cheng, and Yu-Jen Chen. "Computer-aided classification of lung nodules on computed tomography images via deep learning technique." *OncoTargets and therapy* 8 (2015).
- [5] Lakhani, Paras, and Baskaran Sundaram. "Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks." *Radiology* 284, no. 2 (2017): 574-582.
- [6] Ozdemir, Onur, Rebecca L. Russell, and Andrew A. Berlin. "A 3D Probabilistic Deep Learning System for Detection and Diagnosis of Lung Cancer Using Low-Dose CT Scans." *arXiv preprint arXiv:1902.03233* (2019).
- [7] McKinney, Scott Mayer, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back et al. "International evaluation of an AI system for breast cancer screening." *Nature* 577, no. 7788 (2020): 89-94.
- [8] Cho, Junghwan, Kyewook Lee, Ellie Shin, Garry Choy, and Synho Do. "How much data is needed to train a medical image deep learning system to achieve necessary high accuracy?." *arXiv preprint arXiv:1511.06348* (2015).
- [9] Siegel, R.L., Miller, K.D. and Jemal, A. (2020), Cancer statistics, 2020. *CA A Cancer J Clin*, 70: 7-30. doi:10.3322/caac.21590
- [10] Menon, S. , Chapman, D. , Nguyen, P. , Yesha, Y. , Morris, M. , Saboury, B. (2019) "Deep Expectation-Maximization for Semi-Supervised Lung Cancer Screening", Proceedings of ACM SIGKDD 2019, Anchorage, Alaska.
- [11] Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. "Maximum likelihood from incomplete data via the EM algorithm." *Journal of the Royal Statistical Society: Series B (Methodological)* 39, no. 1 (1977): 1-22.
- [12] Zhu, Xiaojin Jerry. *Semi-supervised learning literature survey*. University of Wisconsin-Madison Department of Computer Sciences, 2005.

- [13] Castelli, Vittorio, and Thomas M. Cover. "The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter." *IEEE Transactions on information theory* 42, no. 6 (1996): 2102-2117.
- [14] Papandreou, George, Liang-Chieh Chen, Kevin P. Murphy, and Alan L. Yuille. "Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation." In *Proceedings of the IEEE international conference on computer vision*, pp. 1742-1750. 2015.
- [15] Lindenbaum, Michael, Shaul Markovitch, and Dmitry Rusakov. "Selective sampling for nearest neighbor classifiers." *Machine learning* 54, no. 2 (2004): 125-152.
- [16] Mahapatra, Dwarikanath, Behzad Bozorgtabar, Jean-Philippe Thiran, and Mauricio Reyes. "Efficient active learning for image classification and segmentation using a sample selection and conditional generative adversarial network." In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 580-588. Springer, Cham, 2018.
- [17] Gal, Yarin, Riashat Islam, and Zoubin Ghahramani. "Deep bayesian active learning with image data." In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1183-1192. JMLR. org, 2017.
- [18] Beluch, William H., Tim Genewein, Andreas Nürnberger, and Jan M. Köhler. "The power of ensembles for active learning in image classification." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9368-9377. 2018.
- [19] Tran, Toan, Thanh-Toan Do, Ian Reid, and Gustavo Carneiro. "Bayesian generative active deep learning." *arXiv preprint arXiv:1904.11643* (2019).
- [20] Padmanabhan, Raghav K., Vinay H. Somasundar, Sandra D. Griffith, Jianliang Zhu, Drew Samoyedny, Kay See Tan, Jiahao Hu et al. "An active learning approach for rapid characterization of endothelial cells in human tumors." *PLoS one* 9, no. 3 (2014).
- [21] Danziger, Samuel A., Roberta Baronio, Lydia Ho, Linda Hall, Kirsty Salmon, G. Wesley Hatfield, Peter Kaiser, and Richard H. Lathrop. "Predicting positive p53 cancer rescue regions using Most Informative Positive (MIP) active learning." *PLoS computational biology* 5, no. 9 (2009).
- [22] Shi, Xueying, Qi Dou, Cheng Xue, Jing Qin, Hao Chen, and Pheng-Ann Heng. "An Active Learning Approach for Reducing Annotation Cost in Skin Lesion Analysis." In *International Workshop on Machine Learning in Medical Imaging*, pp. 628-636. Springer, Cham, 2019.
- [23] Benitez, Matias, James Tian, Mark Kelly, Vignesh Selvakumaran, Matthew Phelan, Maciej Mazurowski, Joseph Y. Lo, Geoffrey D. Rubin, and Ricardo Henao. "Combining deep learning methods and human knowledge to identify abnormalities in computed tomography (CT) reports." In *Medical Imaging 2019: Computer-Aided Diagnosis*, vol. 10950, p. 109500V. International Society for Optics and Photonics, 2019.
- [24] Causey, Jason L., Junyu Zhang, Shiqian Ma, Bo Jiang, Jake A. Qualls, David G. Politte, Fred Prior, Shuzhong Zhang, and Xiuzhen Huang. "Highly accurate model for prediction of lung nodule malignancy with CT scans." *Scientific reports* 8, no. 1 (2018): 1-12.
- [25] McNitt-Gray, M. F., Armato III, S. G., Meyer, C. R., Reeves, A. P., McLennan, G., Pais, R. C., ... & Laderach, G. E. (2007). The Lung Image Database Consortium (LIDC) data collection process for nodule detection and annotation. *Academic radiology*, 14(12), 1464-1474.
- [26] Armato III, S. G., McLennan, G., Bidaut, L., McNitt-Gray, M. F., Meyer, C. R., Reeves, A. P., ... & Kazerooni, E. A. (2011). The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. *Medical physics*, 38(2), 915-931.