Spatial Pyramid Pooling with 3D Convolution Improves Lung Cancer Detection

Jason L. Causey, Keyu Li, Xianghao Chen, Wei Dong, Karl Walker, Jake A. Qualls, Jonathan Stubblefield, Jason H Moore, Yuanfang Guan and Xiuzhen Huang

Abstract—Lung cancer is the leading cause of cancer deaths. Low-dose computed tomography (CT) screening has been shown to significantly reduce lung cancer mortality but suffers from a high false positive rate that leads to unnecessary diagnostic procedures. The development of deep learning techniques has the potential to help improve lung cancer screening technology. Here we present the algorithm, DeepScreener, which can predict a patient's cancer status from a volumetric lung CT scan. DeepScreener is based on our model of Spatial Pyramid Pooling, which ranked 16th of 1972 teams (top 1%) in the Data Science Bowl 2017 competition (DSB2017), evaluated with the challenge datasets. Here we test the algorithm with an independent set of 1449 low-dose CT scans of the National Lung Screening Trial (NLST) cohort, and we find that DeepScreener has consistent performance of high accuracy. Furthermore, by combining Spatial Pyramid Pooling and 3D Convolution, it achieves an AUC of 0.892, surpassing the previous state-of-the-art algorithms using only 3D convolution. The advancement of deep learning algorithms can potentially help improve lung cancer detection with low-dose CT scans.

Index Terms—Lung cancer screening, Low-dose CT scan, Deep learning algorithm, Convolutional neural network (CNN), Medical imaging

1 Introduction

Lung cancer is the leading cause of cancer deaths and the second most common cancer in both men and women in the United States [1]. Since lung cancer is most often diagnosed at an advanced stage, the overall 5-year survival is poor (at 18%). Therefore, early detection is the key to improve survival by intervention. Compared with radiographs, low-dose CT can provide more detailed information and has been reported to lead to a 20% lower mortality rate [2]. Low-dose CT has been recommended for lung cancer screening by the US Preventive Services Task Force [3].

Traditional lung cancer screening studies based on examination by human experts have reported false-positive rates as high as 58% [4]. A high false-positive rate not only increases the cost of further tests and surgical procedures but also causes unnecessary anxiety for patients and their families. The development of powerful computer-aided approaches for early lung cancer screening is critical to improve the current clinical practice of CT imaging assessment. Computer-aided approaches aim to produce automated solutions for early lung cancer screening and a reduced false positive rate in diagnosis.

Numerous computer-aided approaches have been developed for chest image analysis in the past fifty years. Ginneken [5] reviews computer analysis in chest imaging and illustrates how the three types of approaches — rule-based image processing, machine learning, and deep learning — have been applied. Moreover, the article showed how deep learning is currently becoming the dominant approach with very promising results [6]. Most computational approaches to date focus on finding and analyzing nodules in lung CT images [7], [8], [9], [10], [11]. However, depending on predefined objects of interest requires detection and segmentation steps that are difficult to automate and limit the applicability of these approaches for automated screening.

Here in this paper, we present the algorithm DeepScreener, which can predict a patient's cancer status from a volumetric lung CT scan. The algorithm is based on the model of Spatial Pyramid Pooling, which we developed in 2017 for the Data Science Bowl (DSB) 2017 competition, with a final rank 16th of 1972 teams (top 1%). The Spatial Pyramid Pooling model employed a pseudo-3-D model considering context information of consecutive slices of a participant's lungs. Here we evaluate the model's ability to generalize beyond the DSB competition datasets by applying it to an independent cohort drawn from the National Lung Screening Trial (NLST) [2]. We find that the model has consistent performance on the independent cohort. Furthermore, when this model is combined with the model of 3D convolution, it achieves an improved AUC of 0.892,

- *Jason L. Causey, Jake A. Qualls, Jonathan Stubblefield, Xiuzhen Huang are with Department of Computer Science and Molecular Bioscicneces Program, Center for No-Boundary Thinking (CNBT), Arkansas State University, Jonesboro, Arkansas 72467.*
- *Keyu Li, Xianghao Chen, Yuanfang Guan are with Department of Computational Medicine & Bioinformatics, University of Michigan, Ann Arbor, Michigan 48109.*
- *Wei Dong is with Ann Arbor Algorithm, Ann Arbor, Michigan 48103, United States of America*
- *Karl Walker is with Department of Mathematics and Computer Science, University of Arkansas at Pine Bluff, Pine Bluff, Arkansas 55455*
- *Jason H. Moore is with Institute for Biomedical Informatics, University of Pennsylvania, Philadelphia, PA 19104*
- *The three authors, J. Causey, K. Li, and X. Chen, are considered as joint first authors.*
- *Correspondence to: Y. Guan (gyuanfan@umich.edu), X. Huang (xhuang@astate.edu).*

surpassing the previous state-of-the-art approach using only 3D convolution.

2 METHODS

For this study, regarding the National Lung Screening Trial (NLST) dataset (https://cdas.cancer.gov/nlst/), data collection has ended, and information is complete through December 31, 2009; NLST has the ClinicalTrials.gov registration number NCT00047385 (Refer to: https://clinicaltrials.gov/ct2/show/NCT00047385). Ethics approval on the study using the NLST dataset was through National Cancer Institute. For this study, regarding the code development, we followed the rules of the Data Science Bowl 2017 (Refer to: https://www.kaggle.com/c/data-science-bowl-2017/rules).

2.1 Datasets for training, validation, and testing

We carefully selected and prepared the datasets for training, validating and testing our algorithm. We used the following datasets: (1) the LIDC/IDRI cohort data, (2) the LUNA16 Challenge data, (3) the DSB2017 Competition data, and (4) a subset of the NLST cohort data. IRB approval is not required for using these datasets, and permission was granted from NCI to access the NLST data.

The LIDC/IDRI data, LUNA16 data and DSB2017 Competition data have been previously used for various biomedical imaging studies and computational approach development and testing. The NLST data is NCI-controlled data; different research groups get their own permission to use the NLST data set/subset for their study. Please refer to the NCI website for the list of publications related to the NLST cohort (https://biometry.nci.nih.gov/cdas/publications/?study=nlst).

The Lung Image Database Consortium image collection (LIDC/IDRI) [12] consists of diagnostic CT data sets with annotated lesions for 1018 participants. Each study includes images from a clinical thoracic CT scan and an associated XML file that records the results of a two-phase image annotation process performed by four experienced thoracic radiologists. The radiologists annotated each scan by marking regions of interest in three classes: "nodule ≥ 3mm", "nodule < 3mm", and "non-nodule". Each nodule in the "nodule ≥ 3mm" class was then given a malignancy score and a detailed segmentation. The LUNA16 Challenge [13] released a list of additional nodules, which were missed by expert readers who originally annotated the LIDC/IDRI data.

The Kaggle Data Science Bowl 2017 [14] dataset is comprised of 2101 chest CT studies. Among them, 1595 were initially released in stage I of the challenge, with 1397 belonging to the training set and 198 belonging to the testing set. The remaining 506 were released in stage II as a final testing set. Each CT study was labeled as "with cancer" if the associated patient was diagnosed with cancer within one year of the scan, and 'without cancer' otherwise. Crucially, the location or size of nodules is not labeled. This data was partially drawn from the NLST cohort. Care was taken in selecting our test cohort to be independent, as explained below.

We tested the performance of the algorithm using data from the National Lung Screening Trial (NLST). 1663 screens with 1000 negative screens and 663 positive screens were originally selected for this current study. The ground truth labels for each study were defined as the presence or absence of a cancer diagnosis during the NLST trial period [2]. We eliminated 5 screens due to missing image data at the point in time where the screen was marked as "positive". An additional 209 screens were eliminated due to an overlap with training data from DSB2017 (202), LIDC/IDRI (3), or both (4). To identify overlapping images, we applied an image fingerprinting method based on comparing intensity histograms of selected slices from each scan in all three primary source cohorts (LIDC/IDRI, DSB2017, NLST). LUNA16 is a subset of LIDC/IDRI. Fingerprints for individual slices were produced by loading the pixel values from the DICOM image and transforming the pixel intensities to Hounsfield Units (HU). Then an intensity histogram containing 20 bins that are roughly centered in regions representing different tissue densities was generated. Bin boundaries were fixed to the following HU values: [-1024, -500, -300, -150, -125, -100, -80, -40, -20, 0, 20, 40, 60, 80, 100, 125, 150, 300, 500, 1024, 2048]. Histograms were calculated in this way for each of the first and last ten slices of each scan, ordered by the Instance Number DICOM attribute. The histograms were combined into a fingerprint vector that was utilized for comparing the mean squared error (MSE) of all possible combinations of images from each dataset. This method was chosen to be both relatively computationally efficient and robust against possible changes made to images when migrating from their original datasets into the competition cohort (such as resampling voxel dimensions, reversing the superior/inferior axis, or missing/duplicated slices). We found no evidence of such changes; all overlaps we discovered had an MSE < 0.001, with a large gap between matches and non-matches (MSE > 200). The lowest scoring non-matches were examined visually to confirm that they were not modified versions of the same scans.

The remaining 1449 images consisted of chest CT images with spacing along the superior/inferior axis ranging from 0.0 - 390.0mm (mean: 1.781mm). This wide range is due to a small number of defects in some images (discussed in more detail later). Ignoring images where these defects were present, the slice thickness range is 0.625 – 10.0mm. (mean: 1.771mm). Spacing along the 2-d (x,y) axis of each slice (lateral/medial, anterior/posterior) is in the range 0.480 - 0.977mm (mean: 0.662mm). Images were captured with X-ray peak tube voltage (kVp) in the range 120.0 - 140.0kV (mean: 121.58kV).

2.2 DeepScreener: a novel algorithm to predict lung cancer with low-dose CT scans

DeepScreener provides an automated solution to predict whether a patient has lung cancer based on a low-dose screening CT scan. Please refer to Figure 1 for an overview of the framework of the deep learning algorithm and the training and validation workflow.

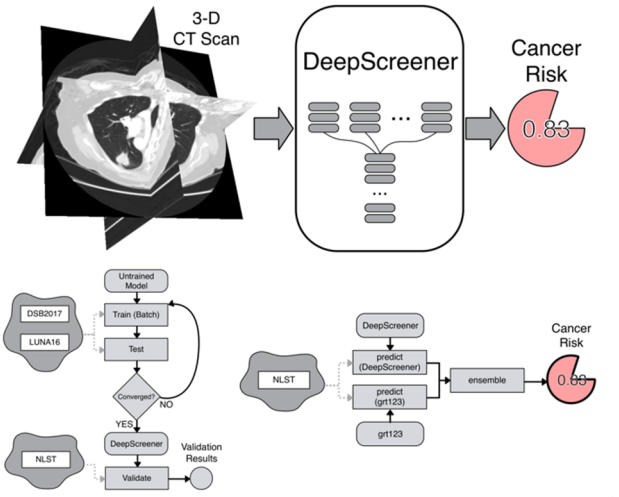


Fig. 1. DeepScreener (upper half) takes a 3-D chest CT image as input and uses a classification model based on convolutional neural networks and gradient boosting decision trees to output a prediction (in the range [0,1]) representing the likelihood that the patient has lung cancer. Training and validation workflow is shown in lower-left flow chart. Lower-right shows the structure of the ensemble model.

Our training and testing strategy can be summarized as follows. First, we trained the image analysis stage of the model for nodule identification using the images and radiologist annotations for nodules in the LIDC/IDRI cohort. We also included annotations for additional nodules released by the LUNA16 Challenge. Then we trained the classification stage of the model for predicting the probability of the presence of lung cancer given the input CT image without a-priori nodule annotations. This probability is translated to a binary label by comparing with a threshold, which was chosen as 0.5 for our initial training/testing. For this purpose, we used the DSB2017 Competition [14] stage one CT data for training and stage II CT data for validation. The classification stage was trained by minimizing log-loss with respect to the ground truth classifications provided by the DSB2017 competition. Finally, the generalization testing of the algorithm was conducted with the selected cohort of low-dose CT images from the NLST study and the results reported here.

The algorithm, DeepScreener, is based on our model of Spatial Pyramid Pooling. We developed the model in 2017 in the DSB2017 competition, which performed well when evaluated using the competition datasets. In the following, we describe the key technical parts of the model in detail.

2.3 Our model of Spatial Pyramid Pooling

The model of Spatial Pyramid Pooling uses consecutive slices and multi-task features to determine whether a nodule is likely to be cancer, and a spatial pyramid to detect nodules at different scales.

Pseudo 3-D Model to Extract Consecutive Information across Slices. We considered a CT scan as a 3-dimensional volume. For example, a typical chest CT scan is about 512 x 512 x N, where N is the number of slices. The resolution within the 2-D slices may be different than between slices. A standard convolutional network can only handle 2-D data, and the

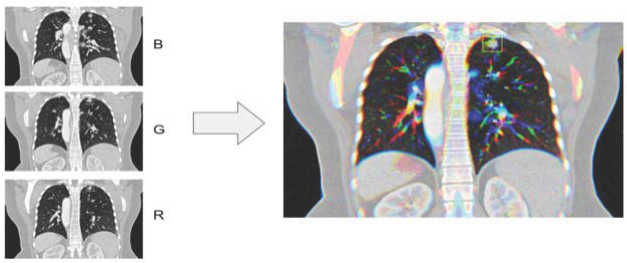


Fig. 2. Pseudo-3-D image produced by "stacking" three individual slices from the 3-D CT image into the blue (B), green (G), and red (R) color channels of a 2-D 3-channel image.

scan has to be processed as N individual slices of 512 x 512. Such slice-based processing loses almost all contextual information along the third dimension. For example, a blood vessel facing the z-dimension (perpendicular to the image in an axial view orientation) appears as a sphere and might be mistaken treated as a small nodule. Notice that a 3-D convolutional network could be used to handle the 3-D information, but a 3-D network its limitations. For example, compared with a 2-D model, a 3-D convolutional network has many more parameters and thus more difficult to train. Training a 3-D network typically requires a much larger training data set. Instead, we chose to use a pseudo-3-D model. Our approach takes advantage of the fact that an image can have multiple (typically 3) channels and encode neighboring slices as multiple channels of a single image. Specifically, for each slice processed, we use the slice itself as the "green" channel of the image and add one slice above as the "blue" channel and once slice below as the "red" channel, each at a distance of 4mm; see Figure 2.

Multi-Task Learning for Feature Extraction for Cancer Classification of the Detected Nodules. A segmentation network [15, 17] only produces a 2-dimensional shape for each nodule detected, and the shape boundary is typically blurry due to low decision confidence. It is possible to extract a few features, like area, average confidence and aspect ratio, but such features extracted solely based on a 2-D shape cannot capture all the characteristics of a nodule that are visible to an expert viewing the original volumetric image.

The LIDC/IDRI dataset [2] provides expert annotation of about 1000 CT scans. In addition to nodule contours, a series of descriptive features are provided for each nodule, e.g. subtlety, sphericity, lobulation, etc. We designed a multi-task convolutional network to simultaneously fit 9 such features (see Figure 3): subtlety, sphericity, margin, lobulation, spiculation, texture, malignancy, calcification-1 and calcification -2. We did not use all of the available categorical features provided by LIDC/IDRI because we found some features are redundant. We split the categorical feature "calcification" into two binary features. The feature extraction network was trained using the LIDC/IDRI annotations as ground truth, with the goal of producing the same numeric ratings for each of these features. The feature extraction network can further increase the information available to the subsequent machine-learning module, i.e. gradient boosting decision trees (GBDT) [18] and improve classification stage accuracy.

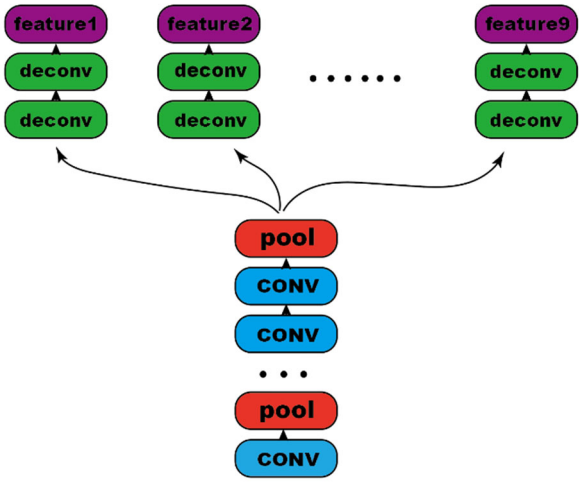


Fig. 3. Multi-task convolutional network simultaneously computes 9 features corresponding to those annotated in the LIDC/IDRI cohort: subtlety, sphericity, margin, lobulation, spiculation, texture, malignancy, calcification-1 and calcification-2.

Feature Pooling with Spatial Pyramid to Detect Tumors of Different Scales. After segmentation, nodule detection and feature extraction, we converted each CT volume into a list of nodule location (x, y, z) and features (size, subtlety, etc). For subsequent learning with GBDT, we pooled this list of variable lengths into a vector of a fixed number of dimensions. We applied the spatial pyramid approach [19] for such pooling. Specifically, we defined a fixed number of regions with overlap, by partitioning the 3-D volume in multiple ways. The image in Figure 4 shows two sample partitions, each with four regions. For each region, we used the feature vector of the largest nodule as the region feature vector, or zeros if no nodule is detected within this region. We then concatenated the feature vectors from all regions to produce a feature vector representing the full CT volume. Even though the spatial pyramid generated a holistic representation of a full CT scan, we can also use it to represent an individual nodule, simply by removing all other nodules detected from the same CT scan. In this way, we can apply the GBDT classifier model to assign a confidence score for each nodule. The patient-level classifier utilizes this ensemble of scores to produce a single confidence score in the range [0,1] which is translated to the binary "cancer" or "no cancer" label by thresholding (default threshold: 0.5).

2.4 The 3D Convolutional Model of *grt123*

The 3D Convolution model we consider is the algorithm *grt123*, which is the winning algorithm of the DSB2017 competition. This 3D convolutional neural network is a unified framework of lung nodule detection and cancer classification. For lung nodule detection, the network is composed of five groups of 3D residual blocks interleaved with four pooling layers, and a set of lateral and feedback connections. For cancer classification, the model selects top five proposals based on the confidence score in the detection network. The model extracts the last convolutional layer of the detection network for each proposal, which is a 32 by 32 by 32 cube of 128 features. For the details of the 3D Convolutional model of the algorithm *grt123*, refer to: https://github.com/lfz/DSB2017/blob/master/solution-grt123-team.pdf.

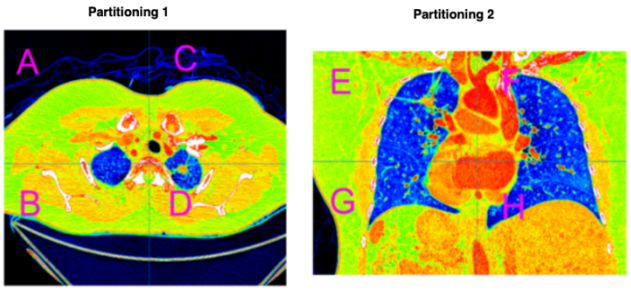


Fig. 4. Example of two partitioning schemes: Each partitioning scheme defines four regions, which may overlap.

3 Evaluation Metrics

The following statistical criteria are used to test the performance of the different models: Accuracy, sensitivity, specificity, AUC, the f1-score and LogLoss.

Accuracy is defined as acc=(TP+TN)/N, where TP represents the number of true positives, TN represents the number of true negatives, and N represents the total number of scans considered. Sensitivity is defined as sen=TP/(TP+FN), where TP is the number of true positives and FN is the total number of false-negative (i.e. missed positive) scans. Specificity is defined as spc=TN/(TN+FP), where TN is the number of true negatives and FP is the number of false-positive scans. We use AUC to refer to the area under the receiver operating characteristic curve, which plots the true-positive rate against the false-positive rate under varying classification threshold values [22]. We used AUPRC to refer to the area under the Precision-Recall curve, which plots the trade-off between precision and recall (recall is synonymous with sensitivity).

The f1-score is a measure of accuracy involving both precision and sensitivity [23], defined as $F_1 = 2 \cdot \frac{Precision \cdot Sensitivity}{Precision + Sensitivity}$ where $Precision = \frac{TP}{TP+FP}$. Log-Loss is defined as $LogLoss = -\frac{1}{N}\sum_{i=1}^{N} [y_i \cdot log_e(\hat{y}_i) + (1 - y_i) \cdot log_e(1 - \hat{y}_i)]$, where $\hat{y}_i$ is the predicted probability of the image belonging to a patient with cancer, y_i is 1 if the diagnosis is cancer, 0 otherwise, $log_e(\cdot)$ is the natural (base e) logarithm.

4 Results

Our algorithm, DeepScreener, can predict a patient's cancer status from a volumetric lung CT scan (refer to Figure 1). We test the performance of our algorithm with the independent set of low-dose CT scans of the NLST cohort.

In the following, we provide the performance details of the three models: our model of Spatial Pyramid Pooling,

the model of 3D Convolution, and the combined model. See Table 1 for a summary of the performance of all models.

4.1 The performance of our model of Spatial Pyramid Pooling on the NLST cohort

We tested the performance of our model of Spatial Pyramid Pooling using 1449 low-dose CT studies obtained from Cancer Imaging Archive [16] with permission from the National Cancer Institute. The model was able to make predictions for 1359 of 1449 CT scans with an accuracy of 78.2%, AUC of 0.858, the area under Precision-Recall curve (AUPRC) of 0.788, and log-loss of 0.484. Refer to Figure 5 for the ROC curve and the Precision-Recall curve resulting from this analysis. The model correctly identified 148 of 432 positive examples (sensitivity 34.3%) and 915 of 927 negative examples (specificity 98.7%). Refer to the Discussions section for an exploration of why the sensitivity/specificity are imbalanced for both our model and grt123, and potential remediations.

We found that several of the images in our test cohort exhibited some kind of unusual defect. For example, some images contained uneven "slice spacing" — the spacing between slices along the superior/inferior axis was not consistent given the spacing information from the image's associated metadata. If the spacing varied enough to suggest a "gap" or missing slice, the image was rejected. Such instances did not appear to be due to systematic or "purposeful" varying of slice thickness – instead, we believe they represented a data quality issue within the image file structure. We also noticed that some images contained one or more "duplicate" slices — the pixels in the 2-D slice were identical to the pixels in another 2-D slice within the scan. In these cases, we dropped the duplicate slice in preprocessing. In total, our Spatial Pyramid Pooling model rejected 90 CT studies due to inconsistencies.

4.2 The performance of 3D Convolution of the winning algorithm grt123 of DSB2017 on the NLST cohort

As a performance comparison to Spatial Pyramid Pooling, We then test the performance of the 3D Convolution of the winning algorithm grt123 of Data Science Bowl 2017 for lung cancer detection (ref: https://datasciencebowl.com/2017algorithms/). On this NLST cohort of low-dose CT scans, the performance of the model of 3D Convolution is very close to the model of Spatial Pyramid Pooling. The 3D Convolution of the algorithm grt123, was able to process 1449 of the 1449 CT scans with an accuracy of 82.1%, AUC of 0.885, area under the Precision-Recall curve (AUPRC) of 0.837, and log-loss of 0.434. Refer to Figure 6 for the ROC curve and the Precision-Recall curve. The algorithm correctly identified 222 of 469 positive examples (sensitivity 47.3%) and 967 of 980 negative examples (specificity 98.7%). Note that grt123 did not require strict image quality control on input images and did not reject any of the input images.

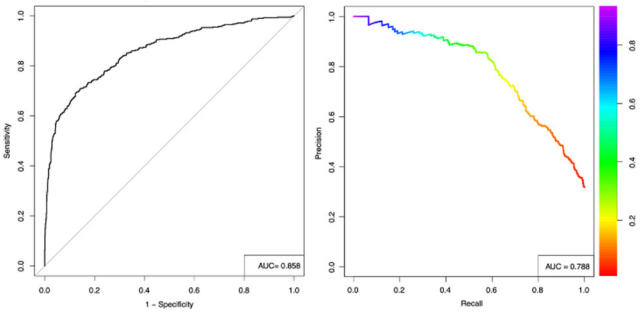


Fig. 5. Performance characteristics for our model of Spatial Pyramid Pooling applied to the selected NLST subset. (a) Receiver operating characteristic curve. Area under the ROC curve is 0.858. (b) Precision-Recall curve. Area under the PR curve is 0.788.

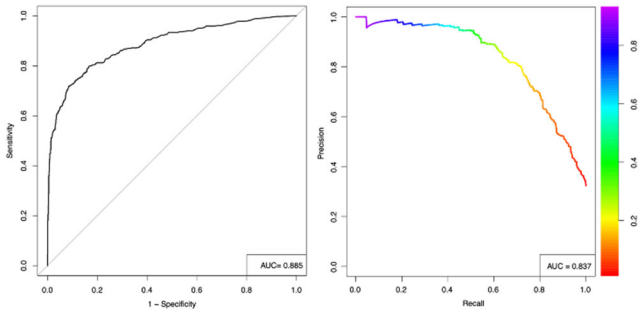


Fig. 6. Performance characteristics for the model of 3D Convolution of the grt123 algorithm applied to the selected NLST subset. (a) Receiver operating characteristic curve. Area under the ROC curve is 0.885. (b) Precision-Recall curve. Area under the PR curve is 0.837.

4.3 The Ensemble Model, with Spatial Pyramid Pooling and 3D Convolution, outperforms the model using only 3D convolution.

We then test the performance of the ensemble model with Spatial Pyramid Pooling and 3D Convolution, using the NLST cohort of low-dose CT scans. See Figure 1 (lower-right) for a diagram of the ensemble architecture. The ensemble model surpasses the performance of the two individual models: Spatial Pyramid Pooling and 3D Convolution; refer to Figure 7. While Spatial Pyramid Pooling has an AUC of 0.858, and 3D Convolution has an AUC of 0.885, the ensemble model (on the 1359 images predicted by both models) achieves an AUC of 0.892 on this NLST cohort, with accuracy of 81.1%, area under the Precision-Recall curve (AUPRC) of 0.848, and log-loss of 0.430. The ensemble correctly identified 183 of 432 positive examples (sensitivity 42.4%) and 919 of 927 negative examples (specificity 99.1%). Through complementing the 3D Convolution model with Spatial Pyramid Pooling, the ensemble model improves lung cancer detection with lung screening low-dose CTs.

We can see that the deep learning models all achieve consistent performance of high accuracy, when tested with the independent set of 1449 low-dose CT scans of the NLST cohort. And when combined the model of Spatial Pyramid Pooling with the model of 3D convolution, it surpasses the previous state-of-the-art approaches using only 3D convolution.

We would like to point out two interesting points from our testing: (1) The performance of these deep learning models, (including our model of Spatial Pyramid Pooling and the 3D Convolution model of algorithm grt123), remained stable with respect to a challenging new dataset (an independent dataset of NLST), indicating that deep learning models can be more broadly applied. (2) The metric used for scoring this particular competition of DSB2017 pushed both the winning algorithm grt123 and our own model toward a particular performance profile that exhibits good detection of "large" cancers at the expense of an undesirably high false-negative rate (seen in results as a high specificity but low sensitivity).

Table 1. Performance metrics for our model of Spatial Pyramid Pooling of DeepScreener, the model of 3D Convolution of grt123, and the ensemble model.

Performance Metric	Spatial Pyramid Pooling	grt123 Model	Ensemble Model
Total	1359	1449	1359
# Positive	432	469	432
# Negative	927	980	927
AUC	0.858	0.885	0.892
AUPRC	0.788	0.837	0.848
Accuracy	0.782	0.821	0.811
LogLoss	0.484	0.434	0.430
f1-score	0.500	0.631	0.587
Sensitivity	0.343	0.473	0.424
Specificity	0.987	0.987	0.991
# False Pos.	12	13	8
# False Neg.	284	247	249

5 Discussions

Computed tomography screening has been shown to aid in early detection of lung cancer in at risk patients, leading to reductions in lung cancer death rates [2]. Unfortunately, CT screening is also associated with high rates of false-positive diagnoses. Currently most computer-aided diagnosis (CAD) tools focus on evaluating lung nodules, which must be identified a-priori, either by a radiologist or with an automated tool. Here we chose to instead focus on risk prediction at the patient level, taking into account information from the whole lung. Our approach could be combined with others to provide a layered strategy for identifying and diagnosing lung cancer. Our approach combines convolutional neural network models to predict the presence of lung cancer at the whole-image level. We chose to test our strategy on low-dose CT scan data from the National Lung Cancer Screening Trial (NLST).

On the NLST cohort of 1449 low-dose CT scans, we tested our deep learning algorithm for predicting lung cancer status with whole low-dose CT scans of the patients. Our algorithm, DeepScreener, was able to make predictions with an AUC of 0.892. From the testing results on the NLST cohort, we anticipate deep learning algorithms can achieve a performance potentially comparable to human experts and radiologists for lung cancer prediction and detection with low-dose CT scans. Through the development of more sophisticated models, as well as training and learning from CT images of even a larger population, deep

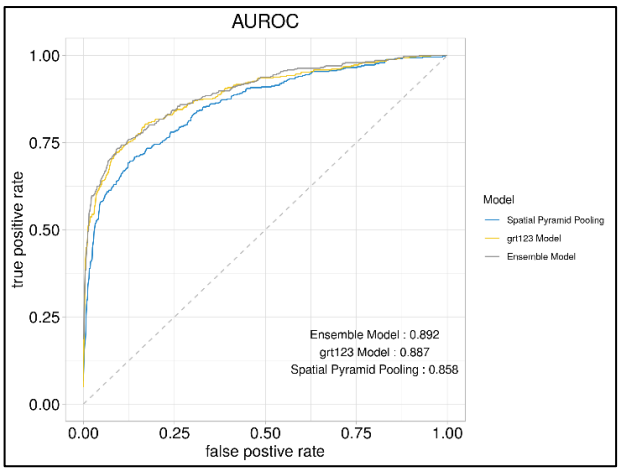


Fig. 7. Performance characteristics for the model of Spatial Pyramid Pooling, the model of 3D Convolution of *grt123*, and the ensemble model, when applied to the NLST cohort. Compared with the model of Spatial Pyramid Pooling, and the 3D Convolution model, the ensemble model achieves an improved AUC of 0.892 on this NLST cohort.

learning algorithms will yield sensitive, stable, consistent and reliable lung cancer screening with the potential of reducing the human effort and cost of screening.

The initial framework of our model of Spatial Pyramid Pooling was developed in 2017 for the DSB2017 competition, and was optimized according to the performance metric used by the competition (minimizing log-loss) [14]. One of our goals for developing automated screening tools is to reduce the false-positive rate associated with lung screenings performed by radiologists. However, the particular metric chosen by the DSB2017 competition may have skewed the model too far in the direction of reducing false positives at the expense of missed cancers (false-negative rate). In addition, the cohort we used for this validation procedure was selected to include cases where the original NLST study contained a likely “false negative” screen. We did this by querying for cases where the patient’s cancer diagnosis followed a negative screen and adding those cases to our query for patients who screened as “positive”. Therefore, our testing cohort itself may be expected to elicit a higher-than-normal false-negative rate. More work needs to be done to balance the trade-off to levels that are clinically acceptable. In evaluating both our competition model and the winning model against a previouslyunseen set of challenging CT screening images, we gained some insight into the value of such competition models in real-world applications. Our results also hint that the choice of a competition scoring metric may induce performance biases in the models that need to be addressed before wider application.

The DSB2017 competition used the LogLoss metric for judging submissions, which is closely related to accuracy in a binary classification problem. As the results show, the accuracy can be high (and LogLoss low) even if the performance is not well balanced between sensitivity and specificity. To explore whether tuning could offset this effect, we examined the effect of modifying the decision threshold (which was at the default 0.5 for the results reported here) with respect to the model's output predictions on NLST and found that a lower threshold would have improved the results. A threshold setting of 0.29 would maximize the accuracy metric at 83.4%, (vs our reported result of 78.2%) and a threshold setting of 0.19 would maximize the sum of sensitivity and specificity at 69.2% and 87.6% respectively, comparatively more balanced than our reported result (34.3% and 98.7%). Obviously, any such tuning based on a-posteriori observations would need to be evaluated by re-training the model and performing an independent validation, but it suggested that training with an objective different than log-loss may be beneficial, as well as parameter tuning to guide the algorithm toward more balanced performance. We would suggest that future competition organizers consider this when choosing a scoring metric, especially when the application is in the medical domain.

In the future, we hope to further develop the model to decrease the number of missed positives and also add visualization options to help make the model's classification decisions interpretable for researchers and clinicians. These improvements will be necessary for tools like these to be accepted into the clinical diagnostic toolset.

Note that there is a recent publication [25], closely related to our work here. In May 2019, Ardila et al. [25] from Google AI published a letter in Nature Medicine, which presented an application of deep learning models on the problem of end-to-end lung cancer screening in the low-dose computed tomography modality. NLST dataset, as well as a proprietary validation set, were used in their article. They conducted reader studies with six experienced radiologists to compare against their algorithmic approach. Their model outperformed the radiologists in their panel on the study without prior imaging, with an 11% reduction in false positives and a 5% reduction in false negatives. The model that they used was essentially a combination of several already known models, adapted to solve different parts of this problem. When evaluating the prediction of localizations, they used the "Hit@N" measure as a metric. This measure suggests that if any of the top N proposed ROIs include any overlap with a ground truth region of interest, the ROI localization is declared a success. It is not as strong a measure as Dice score or Intersection-over-Union, but they do have strong performance by this metric (100% when N=2). The code of their combination model is not available to the public; We cannot conduct the performance comparison of their algorithm and ours.

Data Availibity

The LIDC/IDRI data (https://luna16.grand-challenge.org/data/), LUNA16 data (https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI) and DSB2017 Competition data (https://www.kaggle.com/c/data-science-bowl-2017/data) are publicly available through their individual websites and were previously used for biomedical imaging studies and computational approach development and testing by different research groups in the research field. The NLST data is NCI-controlled data; different research groups get their permission from NCI to use the NLST data for their study. Please refer to the NCI website for the information (https://biometry.nci.nih.gov/cdas/publications/?study=nlst).

Code Availibity

The code is available through Github (https://github.com/aaalgo/plumo), and some intermediate files we processed and generated with this study could be made available to an investigator upon request for academic, research, and noncommercial use.

Acknowledgment

We would like to thank Dr. Fred Prior at University of Arkansas for Medical Sciences (UAMS) for the English proofreading of the preliminary work and for the help with the access to the NLST data.

This research work was partially supported by National Institute of Health NCI grant U01CA187013, and National Science Foundation with grant number 1452211, 1553680, and 1723529, National Institute of Health grant R01LM012601, Arkansas Biosciences Institute grant #200144 "Develop Novel Informatics Algorithms for Lung Cancer Early Screening with CT Scans", as well as was partially supported by National Institute of Health grant from the National Institute of General Medical Sciences (P20GM103429).

References

[1] R. Siegel, K. Miller, and A. Jemal, "Cancer statistics," *CA Cancer Journal of Clinicians* 68, 7-30, 2018.

[2] Aberle, D.R. et al. "Reduced lung-cancer mortality with low-dose computed tomographic screening," *N Engl J Med*. 365, 395-409, 2011.

[3] Lung cancer screening, https://www.uspreventiveservicestaskforce.org/Page/Document/UpdateSummaryFinal/lung-cancer-screening, 2013.

[4] L.S. Kinsinger, et al. "Implementation of lung cancer screening in the veterans health administration," *JAMA Intern Med*. 177, 399–406, 2017. doi:10.1001/jamainternmed.2016.9022

[5] B.V. Ginneken, "Fifty years of computer analysis in chest imaging: rule-based, machine learning, deep learning," *Radiol Phys Technol.* 10, 23–32, 2017.

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature* 521, 436–444, 2015. doi:10.1038/nature14539.

[7] S.C.B. Lo, J.S. Lin, M.T. Freedman, and S.K. Mun, "Computer-assisted diagnosis of lung nodule detection using artificial convolution

neural network," *Proc. SPIE 1898 Medical Imaging,* 1993 *Image Processing,* 1993. doi: 10.1117/12.154572

[8] R. Anirudh, J.J. Thiagarajan, T. Bremer, and H. Kim, "Lung nodule detection using 3d convolutional neural networks trained on weakly labeled data," *SPIE Medical Imaging, International Society for Optics and Photonics*, 978532–978532, 2016.

[9] H.R. Roth, et al. "A new 2.5-d representation for lymph node detection using random sets of deep convolutional neural network observations," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 520–527, 2014.

[10] S. Wang, et al. "Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation," *Medical image analysis*, 40, 172–183, 2017.

[11] J. Causey, et al. "Highly accurate model for prediction of lung nodule malignancy with CT scan," *Scientific Reports* 8, 2018.

[12] 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," *Medical Physics*, 38, 915-931, 2011.

[13] The LUNA16 Challenge. https://luna16.grand-challenge.org/, 2016.

[14] The Data Science Bowl 2017. https://www.kaggle.com/c/data-science-bowl-2017, 2017.

[15] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *CoRR*, abs/1411.4038, 2014.

[16] K. Clark, et al. "The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository," *Journal of digital imaging*, 26, 1045-1057, 2013.

[17] O. Ronneberger, P. Fischer, and B. Thomas, "T. U-Net: Convolutional Networks for Biomedical Image Segmentation," *Medical Image Computing and Computer-Assisted Intervention* (MICCAI), Springer, LNCS, 9351, 234-241, 2015.

[18] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," *Proc. 22nd SIGKDD Conference on Knowledge Discovery and Data Mining*, 2016.

[19] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (CVPR'06), 2169-2178, 2006.

[20] Classification KNN. https://www.mathworks.com/help/stats/classificationknn-class.html, 2019.

[21] P.F. Pinsky, et al. "The National Lung Screening Trial: results stratified by demographics, smoking history, and lung cancer histology," *Cancer*, 119, 3976-83, 2013.

[22] S. J. Mason and N.E. Graham, "Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation," *Quarterly Journal of the Royal Meteorological Society*, 128, 2145–2166, 2002.

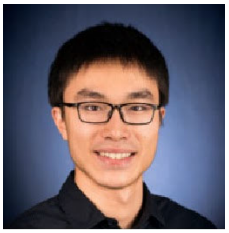
[23] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, 2, 37-63, 2011.

[24] E.R. DeLong, D.M. DeLong and D.L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics*, 837–845, 1988.

[25] D. Ardila, et al. "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," *Nature Medicine*, 25, Letter, 954–961, 2019.

Jason L. Causey is an Assistant Professor of Bioinformatics with the Department of Computer Science, and serves as Associate Director, Center for No-Boundary Thinking (CNBT), Division Lead, CNBT Division of Algorithms and Computational Methodology, Arkansas State University. He received BS degree in Computer Science from Arkansas State University, in 1998, MS degree in Computer Science from Arkansas State University, in 2003, and PhD degree in Bioinformatics from University of Arkansas at Little Rock, in 2017. His research interests include Biomedical Imaging Analysis, Machine Learning on Genomic and Gene Expression Datasets, Machine Learning in Agricultural Contexts, Multi-Modal Biomedical Data Analysis, Enhancing Reproducibility of Research.

Keyu Li is a Master Student in Bioinformatics at the University of Michigan. She got her B.S. degree in Cell & Molecular Biology from McGill University.



Xianghao Chen is a Master Student in Data Science from the University of Michigan. He got his B.S. degree in Environmental Engineering, from Tsinghua University.

Wei Dong is an expert in artificial intelligence, data science, and CTO with IAI, Inc. He received B.S. in Computer Science and Technology and B.S. in Mathematics and Applied Mathematics, Peking University, 2005, and his Ph.D. in Computer Science, Princeton University, 2011.

Karl Walker is Associate Professor and Department Chair, Mathematics and Computer Science, University of Arkansas at Pine Bluff. He received his BS degree in Computer Science from Morehouse College in 2002, MS degree and PhD degree both in Bioinformatics, from University of Arkansas at Little Rock, in 2010 and 2014 respectively.

Jake A. Qualls is an Assistant Professor of Bioinformatics with the Department of Computer Science, and serves as Associate Director, Center for No-Boundary Thinking (CNBT), Division Lead, CNBT Division of Advanced Data Science and Learning, Arkansas State University. He received BS degree in Computer Science from Arkansas State University, in 2002, MS degree in Computer Science from Arkansas State University, in 2004, and PhD degree in Bioinformatics from University of Arkansas at Little Rock, in 2019. His research interests include the application of Machine Learning and Data Science to address cross-disciplinary research questions, Probabilistic reasoning models and their role within Artificial Intelligence, and Computer Science education.

Jonathan Stubblefield is a PhD student, focused on artificial intelligence in biomedical informatics, in the Molecular Biosciences Graduate Program at Arkansas State University. He received his BS in Mathematics and BS in Interdisciplinary Studies from Arkansas State University in 2013, and his MD degree from University of Arkansas for Medical Sciences in 2017.

Jason H Moore is the Edward Rose Professor of Informatics and Director of the Penn Institute for Biomedical Informatics. He also serves as Senior Associate Dean for Informatics and Chief of the Division of Informatics in the Department of Biostatistics, Epidemiology, and Informatics. He received his B.S. (Biological Sciences) Florida State University in 1991, M.S. (Human Genetics) University Of Michigan in 1998, M.A. (Applied Statistics) University Of Michigan in 1998, and Ph.D. (Human Genetics) University Of Michigan in 1999. His research expertise in Artificial intelligence, bioinformatics, biomedical informatics, complex adaptive systems, data science, epistasis, genetic architecture, genetic epidemiology, genomics, human genetics, machine learning, network science, precision medicine, simulation, systems biology, translational bioinformatics, visualization, visual analytics.

Yuanfang Guan is an Associate Professor, Department of Computational Medicine & Bioinformatics, and Associate Professor, Internal Medicine (Nephrology). She got her B.S. degree in Biology from University of Hong Kong in 2005, and Ph.D. degree in Molecular Biology from Princeton University, in 2010. She has her unique contribution to Open Data Science Challenges and the wide application of her algorithms in precision medicine and drug development. She has contributed over 20 best-performing algorithms in the challenges in the community, i.e., DREAM challenges, as well as reaching other communities, such as PhysioNet, Data Science Bowl, and National Data Science Challenge.

Xiuzhen Huang is Professor of Department of Computer Science, and serves as Director of Center for No-Boundary Thinking (CNBT), Arkansas State University. Her work is in the interdisciplinary areas of bioinformatics, biomedical informatics, artificial intelligence, advanced data science, and theory of computation. Dr. Huang conceived and defined the concept of No-Boundary Thinking (NBT). Dr. Huang founded the Arkansas Artificial Intelligence (AI) Campus, and founded the Joint Translational Research Lab on the campuses of Arkansas State University and St. Bernards Medical Center's Internal Medicine Residency Program. She is the Arkansas Research Alliance (ARA) Fellow. Huang received her BS and MS in Computer Science from Shandong University in 1996 and 1999 respectively, and PhD in in Computer Science from Texas A&M University at College Station in 2004.