Towards radiologist-level cancer risk assessment in CT lung screening using deep learning

Stojan Trajanovski[1,*,+]**, Dimitrios Mavroeidis**[1,*,+]**, Christine Leon Swisher**[2,*]**,**
Binyam Gebrekidan Gebre[1]**, Bas Veeling**[3,1]**, Rafael Wiemker**[4]**, Tobias Klinder**[4]**,**
Amir Tahmasebi[5]**, Shawn M. Regis**[6]**, Christoph Wald**[6]**, Brady J. McKee**[6]**,**
Heber MacMahon[7]**, and Homer Pien**[5]

[*]The first three authors have contributed equally.
[+]Corresponding authors: {stojan.trajanovski, dimitrios.mavroeidis}@philips.com.
[1]S.T., D.M., and B.G.G. are with Data Science department, Philips Research, 5656 AE Eindhoven, The Netherlands.
[2]C.L.S. is now with Human Longevity, Inc., San Diego, CA 92121, USA. The research was done while C.L.S. was with Philips Research North America, Cambridge, MA 02141, USA.
[3]B.V. is with Machine Learning lab, University of Amsterdam, 1090 GH Amsterdam and with Philips Research, 5656 AE Eindhoven, The Netherlands.
[4]R.W. and T.K. are with Digital Imaging department, Philips Research, 22335 Hamburg, Germany.
[5]A.T. and H.P. are with Philips Research North America, Cambridge, MA 02141, USA.
[6]S.M.R., C.W., and B.J.M. are with Lahey Hospital & Medical Center, Burlington, MA 01805, USA.
[7]H.M. is with is with the Department of Radiology, University of Chicago, Chicago, IL 60637, USA.

Lung cancer is the leading cause of cancer mortality in the US, responsible for more deaths than breast, prostate, colon and pancreas cancer combined. Recently, it has been demonstrated that screening those at high-risk for lung cancer low-dose computed tomography (CT) of the chest can significantly reduce this death rate. The process of evaluating a chest CT scan involves the identification of nodules that are contained within a scan as well as the evaluation of the likelihood that a nodule is malignant based on its imaging characteristics (such as size, growth, and developing solid nature in the absence of growth). This has motivated researchers to develop image analysis research tools, such as nodule detectors and nodule classifiers that can assist radiologists to make accurate assessments of the patient cancer risk.

In this work, we propose a two-stage framework that can assess the lung cancer risk associated with a low-dose chest CT scan. At the first stage, our framework employs a nodule detector; while in the second stage, we use both the image area around the nodules and nodule features as inputs to a neural network that estimates the malignancy risk of the whole CT scan. The proposed approach: (a) has better performance than the PanCan Risk Model, a widely accepted method for cancer malignancy assessment, achieving around 7% better "Area Under Curve" score in two independent datasets we have employed (NLST data and clinical data obtained at Lahey Hospital & Medical Center (LHMC)); (b) has comparable performance to radiologists in estimating cancer risk at patient level; (c) employs a novel multi-instance weakly-labeled approach to train the deep learning network that requires confirmed cancer diagnosis only at the patient level (not at the nodule level); and (d) employs a large number of lung CT scans (more than 8000) from heterogeneous data sources (NLST, LHMC, and Kaggle competition data) to validate and compare model performance. AUC scores for our model, evaluated against confirmed cancer diagnosis, range between 82% to 90%.

Lung cancer is the leading cause of cancer mortality in the US, responsible for more deaths than breast, prostate, colon and pancreas cancer combined[1]. Survival for lung cancer (5-year) is approximately 18.1% (see e.g.[2]), much lower than other cancer types due to the fact that symptoms of this disease usually only become apparent when the cancer is already at an advanced stage. This highlights the importance of early detection that can potentially extend life expectancy. However, early stage lung cancer (stage I) has a five-year survival of 60-75%. Recently, the National Lung Screening Trial (NLST) demonstrated that lung cancer mortality can be reduced by at least 20% using a screening program of high-risk populations with low-dose computed tomography (CT) of the chest annually[3]. The adoption of such lung cancer screening (LCS) programs at a national level would result in a large volume of CT lung screening (CTLS) scans that need to be assessed by radiologists since the high-risk criteria are met by millions of Americans[4]. This highlights the need to develop useful image analysis tools that can assist radiologists to accurately assess the malignancy risk associated with a CTLS scan and make accurate recommendations for cases that require care escalation, e.g. pulmonology, oncology or thoracic surgical evaluation.

In this paper, we propose a two-stage machine learning framework that can estimate the cancer risk associated with a given CTLS scan. The first stage employs a nodule detector[5] that is used to identify the nodules that are contained within a

scan. Nodule detection is a necessary but insufficient step; a second step is used to evaluate whether the identified nodules are malignant or benign. To achieve this, we employ a neural network inspired by the ResNet architecture[6], regularized using dropout[7] that performs the cancer risk assessment of the whole CTLS scan.

We summarize our contributions with respect to *Model Performance*, *Machine Learning methodology* and *Model Robustness*:

- *Model Performance*: our framework is shown to have better performance of around 7% "Area Under Curve" (AUC), than a widely-accepted malignancy risk model (*PanCan Risk Model*[8]), and comparable performance to radiologists in assessing cancer risk at the patient level.

- *Machine Learning methodology*: our framework is trained using a multi-instance weakly labeled machine learning framework that requires verified cancer diagnosis only at the patient level and not at the nodule level. This is useful because it enables training of lung cancer models using retrospective studies where the cancer diagnosis information is available only at the patient level which is commonly the case in currently available datasets such as the NLST.

- *Model Robustness*: our framework achieves consistent performance between 82% to 90% AUC score as evaluated against verified cancer diagnosis, across three large (∼8000 volumes) & heterogeneous CTLS datasets (NLST, LHMC and Kaggle competition data).

In what follows, we will summarize existing relevant literature with regard to the main components of our framework: nodule detection and nodule malignancy estimation. Initial research results on automated nodule detection were published almost 3 decades ago[9,10], with most of the efforts focusing on chest-radiography and Computer Tomography (CT) images of the chest. More recent works[11,12] employ machine learning methods such as combinations of several classifiers[13] or k-nearest-neighbor classification[14]. Similar to other domains in medical image analysis, nodule detection has been affected by the developments in Deep Learning and in a recent data challenge (Lung Nodule Analysis *LUNA16*[15]), deep learning algorithms have been almost exclusively used to achieve the top-performances. For a survey of works in this area, one can refer to a recent survey paper on the general topic of deep learning methods for medical image analysis by *Litjens et al.*[16].

In the context of this work we employ a Support Vector Machine (SVM)-based nodule detector that has been previously published[5]. As our empirical results across heterogeneous CTLS datasets demonstrate, this non-deep learning nodule detector achieves better performance than the *PanCan Risk Model*[8] and comparable performance to radiologists. This illustrates the utility of evaluating nodule detectors beyond the simple nodule detection task to understand how effective they can be as a first step in cancer risk assessment (i.e. in some cases performance differences may be attributed to small nodules that are benign and do not affect the malignancy risk when assessed by the CTLS scan).

On the topic of nodule malignancy estimation, several recent works rely (at least partially) on deep learning, e.g.[17–25]. Several papers have used 2D and 3D convolutional networks to detect nodules[20,26–32]. The majority of these studies evaluate the malignancy at the nodule level, rather the patient (full volume) level which is considered in this paper.

McWilliams *et al.*[8] have proposed the *PanCan Risk Model* that uses 9 image and patient features in order to estimate the cancer malignancy risk of a nodule. Moreover, a recent study[33] has shown that the *PanCan Risk Model* performs better than Lung-RADS^TM[34] and the National Comprehensive Cancer Network (NCCN) Lung Cancer Screening Guidelines version 1.2016[35], hence highlighting it as a useful benchmark model for our experimental validation.

During the last few years, several LCS related competitions received overwhelming attention from the scientific community and machine learning enthusiasts, such as ANODE[36,37], LUNA16[15,38] and the National Data Science Bowl 2017[39]. In ANODE, the winning teams relied on hand-crafted features and traditional machine learning, while in the most recent ones (LUNA16 and Kaggle) deep learning approaches were dominant. In particular, the National Data Science Bowl 2017 had approximately 2,000 teams participating some of them with multiple team members totaling around 10,000 individual competitors. In the winning entry[40], the authors proposed a 3D convolutional network for nodule detection, using LUNA16 dataset and additional manual nodule annotations of the Kaggle dataset to train their nodule detector. Subsequently, five detected nodules were used as inputs for the malignancy risk assessment network.

Our approach is different than most of the relevant works in that it is trained using verified cancer diagnosis at the patient level as labels rather than cancer diagnosis at the nodule level. This is useful in two ways. Firstly, it enables training lung cancer models using retrospective studies where the cancer diagnosis information is available only at the patient level, which is commonly the case in relevant datasets such as the NLST. Secondly, it allows for the direct association of the performance of our model to the follow-up exams that needs to be performed to the patient. A false positive outcome for the CTLS scan at the patient level means that there will be an unnecessary care escalation, while a false positive rate assessed only at the nodule level may not allow us to understand how many patients will be affected. We are aware of only one publication that uses only patient level annotations[40]. As compared to this work, our framework is evaluated on radiologist assessments and substantially larger number of cases.

Methods

Data

In order to evaluate the performance of our framework, we use CTLS datasets from several sources: NLST[3], LHMC, Kaggle[39] (from both competition stages) and University of Chicago (UCM) data (NLST subset with radiologist annotations). The main data characteristics are summarized in Table 1. We should note that for the NLST dataset we have used all the diagnosed cancer CTLS scans but only a subset of the benign cases to train and validate our model.

Table 1. Data used in our analysis.

Dataset	Total	positive	train or verif.	nodule annotations	Lung-RADS^{TM} classification	continuous radiologists' scores
		Number of volumes			Metadata	
NLST	3410	680	train (90%) / valid. (10%)	yes	no	no
LHMC	3174	56	valid. only (100%)	yes	yes	no
UCM	197	64	valid. only (100%)	yes	no	yes (for 99 volumes)
Kaggle (stage 1)	1397	362	valid. only (100%)	no	no	no
Kaggle (stage 2)	505	153	valid. only (100%)	no	no	no

We trained our model on a subset of the NLST data[3] (3410 volumes, containing 680 hundred diagnosed cancer cases). The NLST-trained model was subsequently verified on the validation data split of the NLST data and the other datasets. The train *vs* test data split is in the ratio of 90%:10%. When evaluating the model on the validation set of the NLST we made sure to exclude any patients whose scans were used to train the model. The lung cancer screening dataset provided by LHMC contains 3174 CTLS scans (with 56 cancer cases), along with a nodule lexicon table that contains detailed information about the identified nodules (such as size, location, etc.). There is only a small number of cancer cases in LHMC data, but on the other hand, the detailed nodule information allows us to compare our framework with other models from the literature that rely on such nodule-level information[8,34]. Furthermore, UCM hospital has provided additional annotations for 197 volumes of the NLST data (that contain 64 cancer cases), that allow us to compare our model with radiologists' assessment as well as the *PanCan* risk model. When validating our model with UCM data, we trained our neural network using all NLST data excluding only the patients that were included in the UCM study. Finally, we use the data from both stages of a recent lung cancer competition National Data Science Bowl 2017 organized and hosted by Kaggle[39]. We should note that the origin of the Kaggle dataset was not disclosed by the competition organizes therefore we cannot exclude the possibility that our models that have been trained on the NLST may have an overlap with the Kaggle data. This should be taken into account when interpreting the Kaggle results only. In the first stage of the competition, 1397 CTLS volumes were provided (with 362 diagnosed cancer cases), while in the second stage 505 volumes were provided (with 153 cancer cases). The CTLS datasets we have used in our analysis come from heterogeneous sources (different hospitals, image quality, reconstruction filters, etc.) and allowed us to validate the generalization capacity of our framework in the experiments section.

Machine Learning Framework for Cancer Risk Assessment

We propose a two-stage machine/deep learning framework for cancer risk assessment. In the first stage, we employ a nodule detector to identify the nodules that are contained in a CTLS scan while in the second stage we use the ten largest nodules identified by the nodule detector as input to a deep and wide neural network that assesses their cancer risk. The decision to use the ten largest nodules was based on the optimal performance obtained from experiments with different numbers of nodules used as input. The details of the two stages are given in the remainder of this section. The pipeline of the algorithm is shown in Figure 1a.

Nodule detection algorithm. In our framework, we employed an SVM-based nodule detector[5]. The nodule detection pipeline consists of two key steps. At first, multi-thresholding is used for robust detection of an extensive number of nodule candidates that aims to find all true nodules while at the same time keeping candidates at irrelevant locations to a manageable amount. In the second step, the large pool of candidates is systematically reduced via a cascaded SVM. We will provide a high-level description of these two steps in the following paragraphs. Details of this approach have been previously published[5].

Multi-thresholding. Prior to candidate detection, lung segmentation is applied to reduce the volume of interest. The lungs are transformed in a down-sampled volume using a region-growing method and the minimal and maximal positions in x-, y- and z-direction of the segmented volume are finally determined. In the next step, the remaining volume is scanned for potential candidates. The idea is to search in each slice for bright circular or semi-circular objects representing a potential nodule. For that purpose, iso-contours for a broad range of thresholds are determined in each slice. More specifically, the intensity interval from -900 to -300 HU is sampled and for each threshold a binary image and corresponding distance map is computed. Two-dimensional seed points are then created at all ridge points in the distance map. Note that a 2D slice-wise processing was
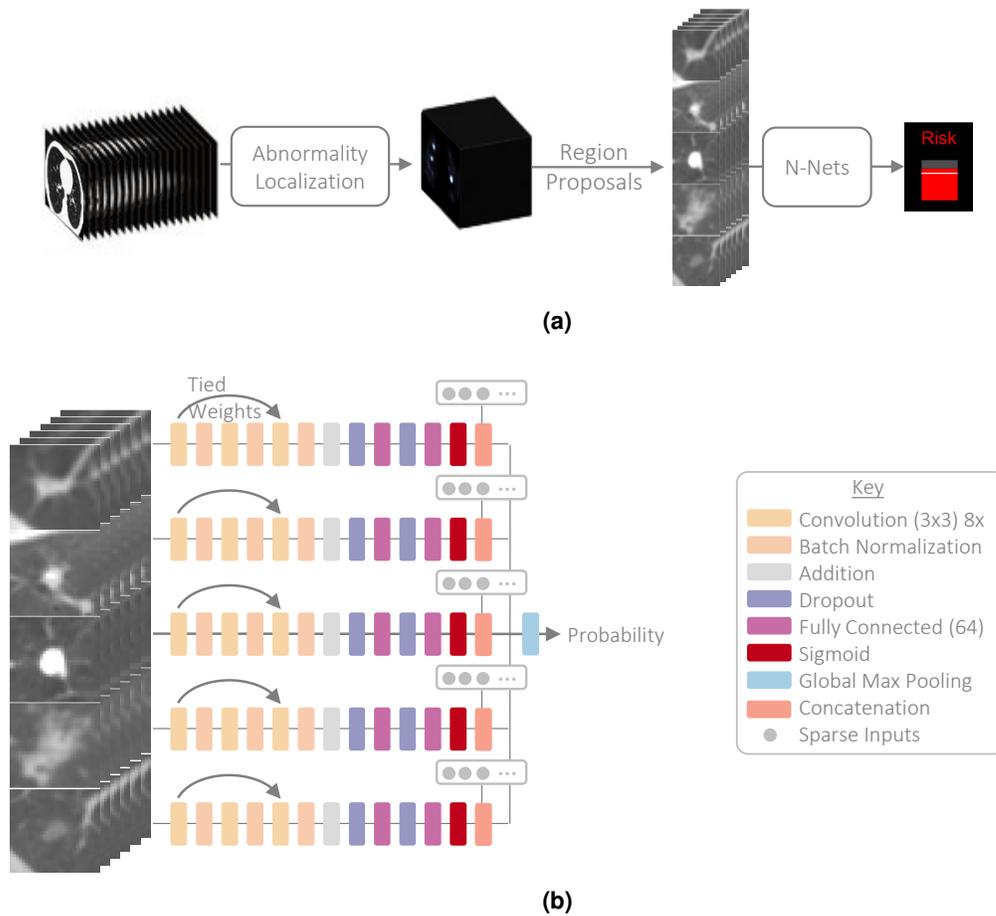
(a)

(b)

Figure 1. Subfigure (a) The pipeline of our algorithm. Initially a nodule detector is used to identify the nodules contained in a CTLS scan. Subsequently, the ten largest nodules are provided as input to a deep learning algorithm that assesses the cancer risk of the scan. Subfigure (b) The network architecture that is applied to the cube around the detected nodules.

preferred over a three-dimensional algorithm to reduce computational complexity and allow for parallel computation. The outcome of this step is an extremely large set of seeds containing many false positives (e.g., placed in vessels, bifurcations, bronchial walls), but ideally containing at least one seed per true nodule. To reduce the number of potential candidate locations, the large set of 2D-seeds is pruned to keep only candidates that belong to 3D-sphere-like objects. The estimation of the object shape is thereby based on the radial-structure tensor[41] determined for multiple 3D iso-surfaces around each seed[5].

Hierarchical SVM. After the first step, we typically obtain more than ten-thousand candidates per volume with usually only a few nodules (if any) per scan. In order to filter out the high number of false positives while also maximizing the true positive rate, a cascaded SVM is employed. For that purpose, a total of 35 image features was used. The extracted characteristics could be generally grouped into geometric features, grayscale features, location features, as well as image properties. From these features, a first SVM was trained that brought down the number of candidates while minimizing the loss of true positives. The second SVM was trained from the remaining candidates. Overall, this approach achieved a sensitivity of 85.9% at 2.5 FP/volume evaluated on the publicly available LIDC/IDRI database.

Deep Neural Network (DNN) for cancer risk assessment. The nodule detector provides us with the nodule locations in all three dimensions: x, y and z as well as additional information such as the nodule size (e.g., radius in mm), the shape of the nodule in terms of the nodule sphericity, and the confidence of the suggestion - given by the svm score. We refer to these parameters as nodule metadata.

(a) Input and augmentation. Based on the output from the previous stage, we can extract from the CTLS scan localized cubes of measuring $32 \times 32 \times 32$ mm^3 around a nodule (and since we employ isotropic resampling to 1 mm^3 each voxel corresponds to 1 mm^3). This gives us sufficient context for the experiments as we find that smaller or larger cubes do not improve and can

even degrade performance. Additionally, during training, a random crop of $28 \times 28 \times 28$ mm^3 out of the extracted $32 \times 32 \times 32$ mm^3 cube is taken to ensure that the network does not see the same images in each batch iteration thus reducing overfitting. Finally, from the 3D $28 \times 28 \times 28$ mm^3 cube we extract three different 2D projections, as channels, namely coronal, sagittal, and transversal, thus ending up with $3 \times 28 \times 28$ input per nodule for the neural network (see Figure 1b left-hand part). Moreover, for each nodule we use additional features, such as nodule radius, sphericity and svm score (confidence level of a detected nodule as provided by the SVM algorithm used by the nodule detector) as numeric inputs added in the penultimate level in the architecture. The nodule descriptors are obtained automatically by the nodule detector without any human intervention. Different volumes have different number of nodules. In the experiments we used the 10 largest nodules, when there are at least 10 nodules in the volume, otherwise all the nodules are used and the remaining "spots" are masked.

(b) Neural network architecture. We use a ResNet-like[6] deep and wide neural network for evaluating the cancer risk associated with each CTLS scan. (Deep refers to the number of layers, while wide refers to the number of inputs.) The input consists of the image part as described in the previous paragraph and the additional nodule features (e.g., radius etc.) of the nodule properties added at the penultimate layer. The network architecture is visualized in Figure 1b. More details of the exact layer configuration of the neural network are given in Table 2 in the supplementary material. We used 3x3 kernels for convolutional neural network blocks with 8 channels, intertwined with batch normalization and additional connections for realizing the ResNet-blocks (see inputs 5. and 6. in Table 2, supplementary material), augmented with dropout for better generalization and followed by fully connected layers (with 64 units) and sigmoid activation functions. Finally, we concatenate the last fully connected layer with the nodule metadata, making the deep neural network also wide. At the end, we perform a global max pooling aggregating over the maximum of ten branches representing the different nodules, which estimates the final cancer risk probability. Interestingly, we can obtain very good performance even in cases when the dropout rate is set to one from $\{0.7, 0.8, 0.9\}$ (i.e. "retaining probability" of around 0.1-0.3), in contrast to what is considered standard practice in the relevant literature[7] (page 1938).

(c) Training of our model and performance evaluation. Our model relies on information about verified cancer diagnosis at the volume/scan level. This implies that our CT volumes were annotated with label 1 in cases where the patient was diagnosed with lung cancer and 0 otherwise. In this sense our data can be categorized as multi-instance weakly labeled, since our labels (cancer diagnosis) are provided for the group of nodules that are contained within a scan and not for each nodule individually. This information was available in all datasets reported in Table 1. Using these labels at the volume level, we trained our neural network with the binary cross-entropy loss function. In the empirical results we always evaluate the performance of our model with respect to verified cancer diagnosis at the volume level.

(d) Alternative network architectures & deep learning experiments. We explored the individual contribution of the architectures attributes with various ablation experiments. Most of the results and how they compare with the proposed model are given in the supplementary material in Table 3 and Figures 5-10. Namely, we tried using small to moderate dropout, using less or more global nodule features (e.g., goodness, brightness, Hounsfield units (HU), *x*, *y* and *z* nodule dimensions), using only a single (largest) nodule, taking larger or smaller part around a nodule, using different architectures such as VGGs[42] and DenseNets[43]. The results suggest that there is no benefit of these architectures and the proposed one (in Table 2 and Figure 1a) performs better than the alternative architectures or hyper-parameters.

(e) Visualizations and Grad-CAMs. To understand which parts of a given image are the main areas used by the network to calculate the cancer risk, we use a visualization technique called Gradient-based Class Activation Mapping (Grad-CAM for short)[44]. The advantage of Grad-CAM over other visualization techniques such as deconvolution[45] or guided backpropagation[46] is that the visualizations from Grad-CAM are class discriminative and can therefore help us understand better the reasoning behind the networks' decisions. It should be noted that for generating the Grad-CAM visualizations, we used a slightly different image input for the DNN algorithm. More precisely, we employed three consecutive axial slices of the detected nodules instead of the sagittal, coronal and transverse slices of the nodule cube. This choice was made because the Grad-CAM algorithm could not differentiate between the three input slices and thus the visualization when using the sagittal, coronal and transverse slices as input were less intuitive.

PanCan Risk Model

To empirically validate our framework, we employ a model developed at the Vancouver General Hospital for nodule malignancy estimation [8]. This method provides information using a single scan, and does not use information potentially available from multiple scans of the patient (that could be used, for example, to identify nodule growth). The model employs a formula, which calculates the malignancy score based on 9 numerical or Boolean input parameters, including three patient features: age of a patient [number], gender of a patient, lung cancer family history [true or false]; one clinical or image-based feature: presence of emphysema [true or false]; one patient specific image-based feature: nodule count (number of nodules) in the CTLS scan [number]; and four nodule specific image-based features: size of a nodule (diameter) - which is longest in-slice axis [a number], type of the nodule [one of nonsolid, part-solid, solid], location of the nodule in the upper lobe [true or false], and

nodule spiculation [true or false].

$$\text{nodule malignancy score} = \frac{1}{1 + e^{-\sum\limits_{i=1}^{9} \text{weight}_i \cdot \text{input}_i}} \tag{1}$$

To compare our model that produces a single risk score for each CTLS scan to the *PanCan Risk Model* that computes a risk score on a per-nodule basis, we set the CTLS scan malignancy score to be derived by the maximum malignancy score of all nodules. In our experiments this provides the best performance results for the *PanCan* risk score (rather than taking the mean, minimum scores etc. of a nodule per study).

Radiologist predictions

To compare our results to radiologist performance, an observer study was conducted at UCM using 99/197 CTLS scans for which radiologists have provided a continuous numeric estimate of the cancer probability in addition to the Lung-RADS^TM score. This subset consists of 20 malignant and 79 benign cases. Each selected case had to have at least one nodule within the range of 6-25 mm. The selection was made in a way to match the distribution of the nodule sizes in the NLST database. The selection was made in a way to match the distribution of the nodule sizes in whole NLST database.

Besides nodule size distribution matching, the selection covered nodule types of all categories except for calcified nodules. Three senior and three junior radiologists from the thoracic imaging department participated in the study. A graphical user interface was designed for the study to capture and demonstrate relevant information to the user such as the three orthogonal views (axial, sagittal, and coronal) of the imaging focused on the slices containing the nodule as well as demographic information such as sex, age, smoking history, and family history of smoking. The user was able to measure the nodule size using the measurement tool provided. After taking all information into account, the radiologist was asked to provide the assessment of the risk for developing lung cancer in terms of a percentage number.
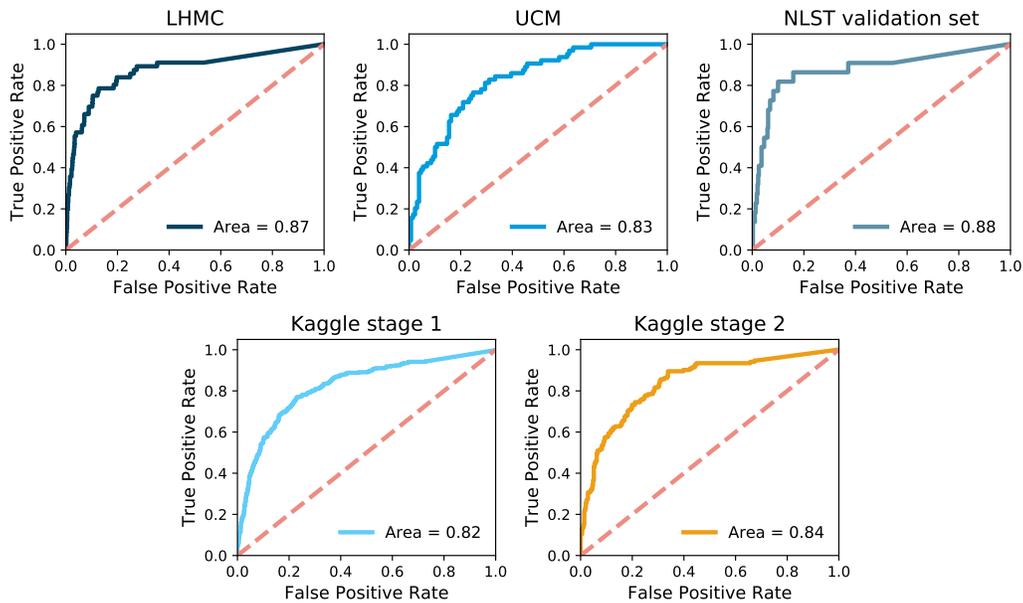

Figure 2. Performance of our model is consistent across different datasets. For the NLST subfigure we have trained the model on the training data split (90% of NLST data) and used the validation split to evaluate the performance making sure that there was no patient overlap between the two splits. We employed the same model (architecture and weights) to generate the LHMC and Kaggle subfigures, trained using the training data-split of NLST with the network architecture described in Table 2 and dropout set to 0.9. For the UCM subfigure, we have used a model trained on all the NLST excluding only the patients that are included in the UCM study the dropout in this experiment was set to 0.8. Experiments with different dropout rates (0.7, 0.8 or 0.9) performed similarly ($\pm$1% AUC).

Results

Performance results

(a) Performance of our algorithm. The performance of our framework was stable across the different datasets used and can achieved an AUC (Area Under the Curve) score between $82\% - 88\%$ as shown in Figure 2. It is worth re-iterating that our model has been trained using only data from one dataset (NLST), but generalizes well across all different datasets that we used in the experiments. Our evaluation is more extensive than the majority of related works that commonly use smaller and less
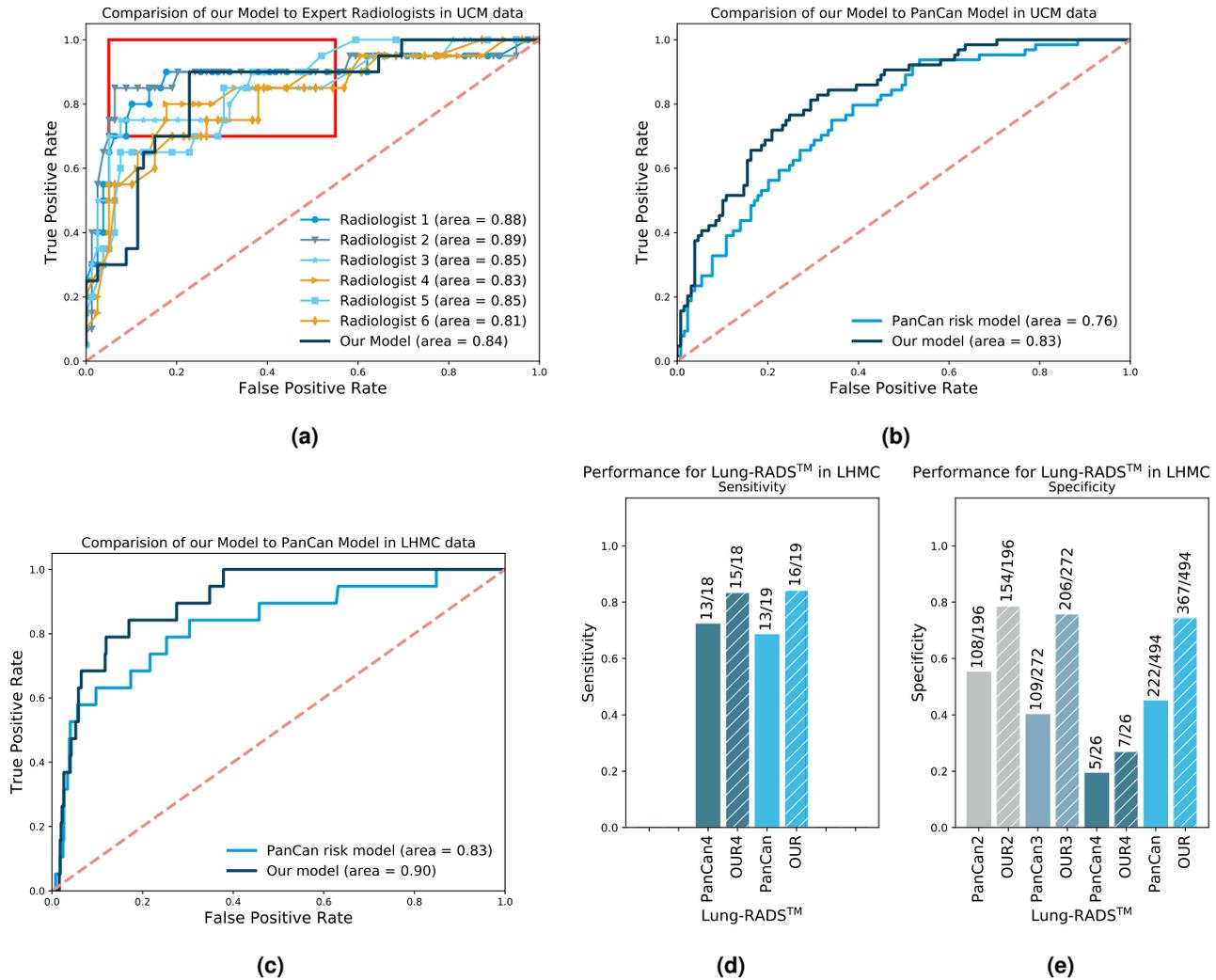

Figure 3. Lung cancer risk assessment performance of our DNN model compared to *PanCan Risk Model*[8] and radiologists for UCM and LHMC data. Figure (a) ROC curve showing the performance of our model compared to radiologists' assessments for 99 studies that have available annotations in UCM data. In this experiment, we have trained our model on the NLST data using the network architecture described in Table 2 with dropout set to 0.7 excluding the patients that were in the UCM study. Figure (b) ROC curve showing the performance of our model and the *PanCan Risk Model* for all 197 studies in UCM data (with 64 verified cancer cases). We have trained our with NLST excluding the patients that were included in the UCM study. Figure (c) ROC curve showing the performance of our model and the *PanCan Risk Model* for LHMC data. Our model have been trained on a 90% train-data split of the NLST. Lung-RADS^TM grouped performance in LHMC data for our framework (labeled as "OUR" and with stripes) and the *PanCan Risk Model* (labeled as "PanCan"): Figure (d) Lung-RADS^TM grouped sensitivity for LHMC data when the specificity is set to 80%; Figure (e) Lung-RADS^TM grouped specificity for LHMC data when the sensitivity is set to 84%. "OUR2,3,4" and "PanCan2,3,4" labels refer to the performance achieved for Lung-RADS^TM = 2,3,4 classified cases, while "PanCan" and "OUR" refer to the full LHMC dataset.

diverse datasets.

(b) Performance of alternative DNN architectures and hyper-parameter choices. We considered different neural network architectures in order to find the optimal one, but also to better understand the effect of the different deep learning hyper-parameters to the problem of lung cancer. The comparison is shown in Table 3 (supplementary material). We can see that although the influence of the largest nodule is large, additional nodules significantly help to boost the performance. Moreover, the additional "wide" inputs, although correlated to the image part can also improve model performance. Aggressive dropout parameter turns out to be very important for generalization rather than moderate or low dropout often used in the literature[7] (page 1938) and we can observe that setting dropout parameter to a high value from $\{0.7, 0.8, 0.9\}$ achieves good performance.

(c) Additional insights from visualizations and Grad-CAMs. Using visualizations and Grad-CAMs, our results demonstrated in Figure 4, show that our DNN model focuses on the nodule surface shape and its margins (spiculation, lobulation, smoothness) and also its proximity to the pleura, something that is also practiced by radiologists. Moreover, when using three consecutive axial slices, the results were more interpretable in comparison to sagittal, coronal and transverse projections as input. We also observed that the algorithm frequently focused on nodule surface that is one the main criteria radiologists use to evaluate malignancy risk.

Comparison with the radiologist performance

Figure 3a shows the ROC curves of our model as compared to the ROC curves obtained by the single-scan risk assessments of the 6 radiologists on the subset of 99 volumes (different patients) of the UCM data out of which 20 correspond to verified cancer cases. Our algorithm shows a comparable and often better performance than the one of radiologists. We highlight with a red box the area of the ROC curve where the true positive rate is at a high level which is an important factor when performing Lung Cancer Screening (i.e. no cancer cases are missed). It should be noted that our work is one of the few studies[24] in the literature where such a comparison to radiologists is performed.

Comparison with the *PanCan Risk Model*

The results, presented in Figures 3b and 3c show that our proposed model significantly outperforms the *PanCan Risk Model*[8] by approximately 7% AUC for both UCM and LHMC datasets. Further, we compare our algorithm with the *PanCan Risk Model* for various Lung-RADS™ categories in the LHMC data. We performed the evaluation by comparing the sensitivity for different fixed specificity performance levels and *vice versa* (i.e. comparing specificity for fixed levels of sensitivity for both algorithms. These evaluation per different Lung-RADS™ categories show that our algorithm performs better than the *PanCan Risk Model* in terms of sensitivity (in Figure 3d) and specificity (in Figure 3e).

Discussion

The topic of lung cancer malignancy risk assessment is an important research topic that has recently attracted a lot of attention due to the fact that there are nearly 10,000,000 people in the US alone fit the high-risk criteria for CTLS. This illustrates the need to develop tools to help radiologists evaluate the CTLS scans and protect the patients without lung cancer from the risks associated with unnecessary care escalation.

In this paper, we propose a two-stage framework for cancer risk assessment that uses a nodule detector for identifying the nodules that are contained in a CTLS scan and subsequently uses the areas around the nodules as input to a neural network that performs the malignancy risk assessment. The algorithm has consistent performance across three different CTLS datasets and it is shown to have better performance than the *PanCan Risk Model*[8]. Moreover, the algorithm has comparable performance to radiologists and also its performance ranks among the top-10 submissions of a recent data challenge related to CTLS[39].

As a focus for further work, one can consider the differences in model performance across different image quality settings such as reconstruction filters (soft-tissue, sharp, etc.). One can potentially improve performance by limiting the neural networks' training and subsequently the prediction on a unique set of reconstruction filters or consider domain adaptation methods to optimize performance across different image quality data.

References

1. Siegel, R., Ma, J., Zou, Z. & Jemal, A. Cancer statistics, 2014. *CA: A Cancer J. for Clin.* **64**, 9–29 (2014).

2. NAACC Review. 2018 state of lung cancer report. https://www.naaccr.org/2018-state-lung-cancer-report/ (2018).

3. The National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. *New Engl. J. Medicine* **365**, 395–409 (2011). DOI 10.1056/NEJMoa1102873. PMID: 21714641.

4. de Koning HJ, R, M., SK, P. & et al. Benefits and harms of computed tomography lung cancer screening strategies: A comparative modeling study for the u.s. preventive services task force. *Annals Intern. Medicine* **160**, 311–320 (2014).
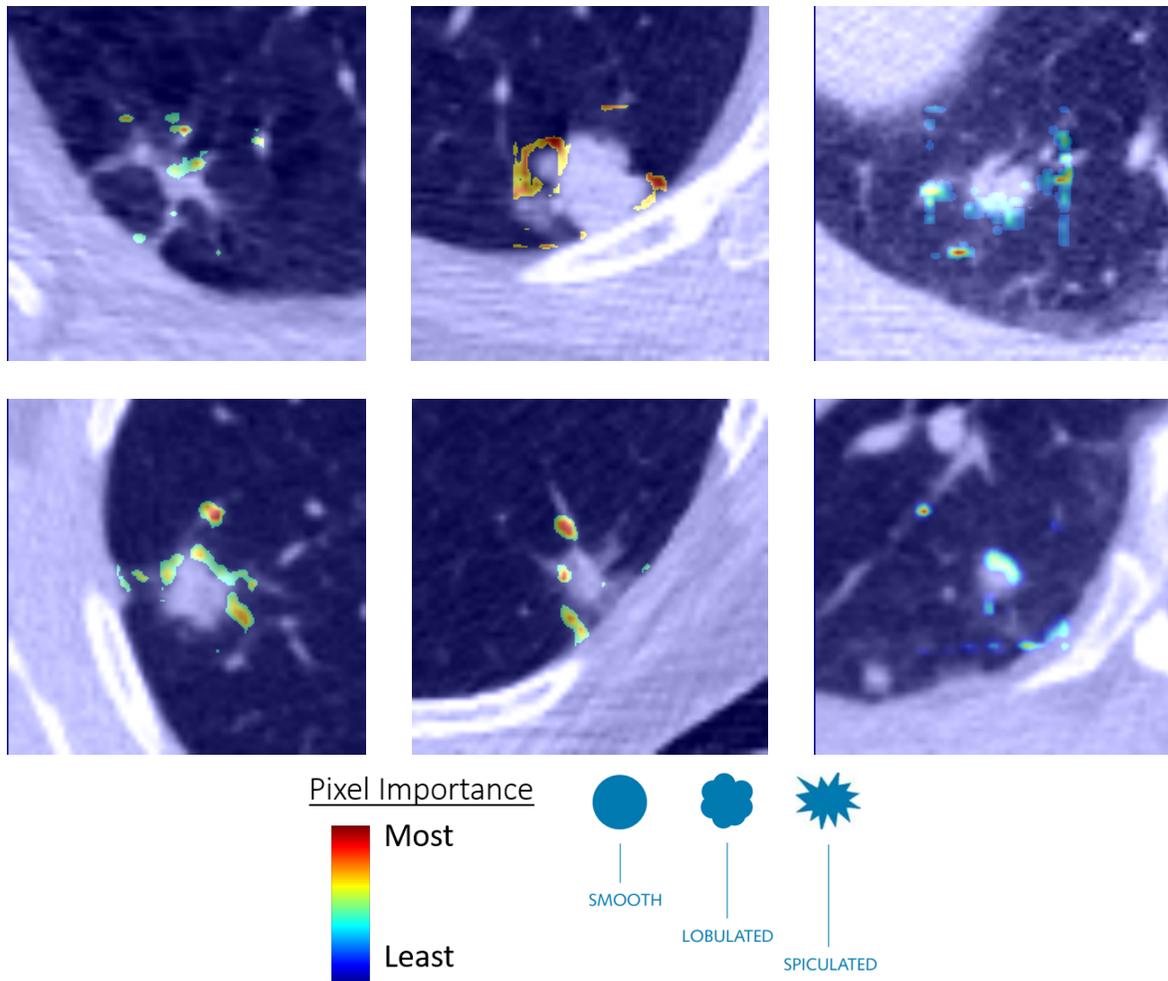
Figure 4. Heatmap generated by the Grad-CAM method overlayed over detected nodules.

5. Bergtholdt, M., Wiemker, R. & Klinder, T. Pulmonary nodule detection using a cascaded svm classifier. In *Proc.SPIE*, vol. 9785, 9785 – 9785 – 11 (2016).

6. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778 (2016).

7. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).

8. McWilliams, A. *et al.* Probability of cancer in pulmonary nodules detected on first screening ct. *New Engl. J. Medicine* **369**, 910–919 (2013). PMID: 24004118.

9. Preteux, F. *A Non-Stationary Markovian Modeling for the Lung Nodule Detection in CT*, 199–204 (Springer Berlin Heidelberg, Berlin, Heidelberg, 1991).

10. Benedict Lo, S., Lin, J., Freedman, M. & Ki Mun, S. Computer-assisted diagnosis of lung nodule detection using artificial convoultion neural network. vol. 1898, 1898 – 1898 – 11 (1993).

11. Messay, T., Hardie, R. C. & Rogers, S. K. A new computationally efficient cad system for pulmonary nodule detection in ct imagery. *Med. Image Analysis* **14**, 390 – 406 (2010).

12. Camarlinghi, N. *et al.* Combination of computer-aided detection algorithms for automatic lung nodule identification. *Int. J. Comput. Assist. Radiol. Surg.* **7**, 455–464 (2012).

13. J. Suárez-Cuenca, J., Guo, W. & Li, Q. Automated detection of pulmonary nodules in ct: False positive reduction by combining multiple classifiers. In *Proc. SPIE Medical Imagings*, vol. 7963 (2011).

14. Murphy, K. *et al.* A large-scale evaluation of automatic pulmonary nodule detection in chest ct using local image features and k-nearest-neighbour classification. *Med. Image Analysis* **13**, 757 – 770 (2009). Includes Special Section on the 12th International Conference on Medical Imaging and Computer Assisted Intervention.

15. Challenge. LUng Nodule Analysis 2016. `https://luna16.grand-challenge.org/` (2018).

16. Litjens, G. *et al.* A survey on deep learning in medical image analysis. *Med. Image Analysis* **42**, 60 – 88 (2017).

17. Ciompi, F. *et al.* Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2d views and a convolutional neural network out-of-the-box. *Med. Image Analysis* **26**, 195 – 202 (2015).

18. van Ginneken, B., A. A. Setio, A., Jacobs, C. & Ciompi, F. Off-the-shelf convolutional neural network features for pulmonary nodule detection in computed tomography scans. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, 286–289 (2015).

19. Hua, K.-L. and Hsu, C.-H. and Chusnul Hidayati, S. and Cheng, W.-H. and Chen, Y.-J. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. *OncoTargets Ther.* **8**, 2015–2022 (2015).

20. Setio, A. A. A. *et al.* Pulmonary nodule detection in ct images: False positive reduction using multi-view convolutional networks. *IEEE Transactions on Med. Imaging* **35**, 1160–1169 (2016).

21. Cheng, J.-Z. *et al.* Computer-aided diagnosis with deep learning architecture: Applications to breast lesions in us images and pulmonary nodules in ct scans. *Sci. Reports* **6** (2016).

22. Shen, W., Zhou, M., Yang, F., Yang, C. & Tian, J. Multi-scale convolutional neural networks for lung nodule classification. In Ourselin, S., Alexander, D. C., Westin, C.-F. & Cardoso, M. J. (eds.) *Information Processing in Medical Imaging: 24th International Conference, IPMI 2015, Sabhal Mor Ostaig, Isle of Skye, UK, June 28 - July 3, 2015, Proceedings*, 588–599 (Springer International Publishing, Cham, 2015).

23. Chen, S. *et al.* Automatic scoring of multiple semantic attributes with multi-task feature leverage: A study on pulmonary nodules in ct images. *IEEE Transactions on Med. Imaging* **36**, 802–814 (2017).

24. van Riel, S. J. *et al.* Malignancy risk estimation of pulmonary nodules in screening cts: Comparison between a computer model and human observers. *PLOS ONE* **12**, 1–15 (2017).

25. Ciompi, F. *et al.* Towards automatic pulmonary nodule management in lung cancer screening with deep learning. *Sci. Reports* **7** (2017).

26. Dou, Q., Chen, H., Yu, L., Qin, J. & Heng, P. A. Multilevel contextual 3-d cnns for false positive reduction in pulmonary nodule detection. *IEEE Transactions on Biomed. Eng.* **64**, 1558–1567 (2017).

27. Shen, W. *et al.* Learning from experts: Developing transferable deep features for patient-level lung cancer prediction", booktitle=”medical image computing and computer-assisted intervention – miccai 2016: 19th international conference, athens, greece, october 17-21, 2016, proceedings, part ii. 124–131 (Springer International Publishing, Cham, 2016).

28. Li, W., Cao, P., Zhao, D. & Wang, J. Pulmonary nodule classification with deep convolutional neural networks on computed tomography images. *Comput. Math. Methods Medicine* **6215085** (2016).

29. Sun, W., Zheng, B. & Qian, W. Computer aided lung cancer diagnosis with deep learning algorithms. vol. 9785, 9785 – 9785 – 8 (2016).

30. Teramoto, A., Fujita, H., Yamamuro, O. & Tamaki, T. Automated detection of pulmonary nodules in pet/ct images: Ensemble false-positive reduction using a convolutional neural network technique. *Med. Phys.* **43**, 2821–2827 (2016).

31. Shin, H. C. *et al.* Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE Transactions on Med. Imaging* **35**, 1285–1298 (2016).

32. Anirudh, R., Thiagarajan, J.-J., Bremer, T. & Kim, H. Lung nodule detection using 3d convolutional neural networks trained on weakly labeled data (2016).

33. van Riel, S. J. *et al.* Malignancy risk estimation of screen-detected nodules at baseline ct: comparison of the pancan model, lung-rads and nccn guidelines. *Eur. Radiol.* **27**, 4019–4029 (2017).

34. Lung-RADS Version 1.0 Assessment Categories. `https://www.acr.org/~/media/ACR/Documents/PDF/QualitySafety/Resources/LungRADS/AssessmentCategories.pdf`. Accessed: 2017-10-25.

35. National Comprehensive Cancer Network (NCCN) Guidelines, Version 1.2016, Lung Cancer Screening, Release date June 23, 2015. `https://www.nccn.org/professionals/physician_gls/f_guidelines.asp#detection`. Accessed: 2017-10-25.

36. Challenge. Automatic Nodule Detection 2009. `https://anode09.grand-challenge.org/` (2009).

37. van Ginneken, B. *et al.* Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: The ANODE09 study. *Med. Image Analysis* **14**, 707 – 722 (2010).

38. Setio, A. A. A. *et al.* Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: The LUNA16 challenge. *Med. Image Analysis* **42**, 1 – 13 (2017).

39. Kaggle competition. Data science bowl 2017: Can you improve lung cancer detection? `https://www.kaggle.com/c/data-science-bowl-2017` (2017).

40. Liao, F., Liang, M., Li, Z., Hu, X. & Song, S. Evaluate the malignancy of pulmonary nodules using the 3d deep leaky noisy-or network. *arXiv preprint arXiv:1711.08324* (2017).

41. Wiemker, R. *et al.* A radial structure tensor and its use for shape-encoding medical visualization of tubular and nodular structures. *IEEE Transactions on Vis. Comput. Graph.* **19**, 353–366 (2013).

42. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *CoRR* **abs/1409.1556** (2014).

43. Huang, G., Liu, Z., van der Maaten, L. & Q. Weinberger, K. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017).

44. Selvaraju, R. R. *et al.* Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization. *arXiv preprint arXiv:1610.02391* (2016).

45. Zeiler, M. D. & Fergus, R. Visualizing and understanding convolutional networks. In *European conference on computer vision*, 818–833 (Springer, 2014).

46. Springenberg, J. T., Dosovitskiy, A., Brox, T. & Riedmiller, M. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806* (2014).

Author contributions statement

S.T., D.M., C.L.S., B.G.G., and B.V. designed and implemented the deep neural networks part of the algorithm. R.W. and T.K. designed and implemented the nodule detector part of the algorithm. S.T., D.M., C.L.S., B.G.G., and B.V. designed, implemented and conducted the experiments and analyzed the results. C.L.S., A.T., S.M.R., C.W., B.J.M. and H.M. participated in the data gathering or provided part of the data. S.T. and D.M. wrote the manuscript. H.P. and D.M. supervised the research. All authors reviewed and approved the manuscript.

Additional information

Competing financial interests. The authors declare no competing financial interests.

Supplementary material

Deep neural network architecture of the proposed model

The deep neural network architecture and how the layers are interconnected is given in Table 2 (also visualized in Figure 1b).

Table 2. Architecture of the deep and wide neural network.

Layer	Properties	Previous layer(s)
1. Image input	(10x3x28x28)	- Img: 10 nod, 3proj 28x28
2. Conv Layer + BN	(3x3, 8x)	1.
3. Conv Layer + BN	(3x3, 8x)	2.
4. Conv Layer + BN	(3x3, 8x)	3.
5. Conv Layer + BN	(3x3, 8x)	1.
6. Addition/Merge + BN	-	4., 5.
7. Dropout + BN	{**0.7, 0.8, 0.9**}	6.
8. Dense + BN	(64)	7.
9. Dropout + BN	{**0.7, 0.8, 0.9**}	8.
10. Dense + BN	(64)	9.
11. Numeric input	(10x1)	- Radius
12. Numeric input	(10x1)	- Sphericity
13. Numeric input	(10x1)	- x, y, z nod. coordinates
14. Numeric input	(10x1)	- svm score
15. Addition/Merge	-	10., 11., 12., 13., 14.
16. Dense + sigmoid	(1)	15.
17. GlobalMaxPool	(10)	16.

Comparison of the best model with other choices of deep learning configurations

In this section, we evaluate the best performing model described in Table 2, in comparison with other choices of deep learning architectures. The results are summarized in Table 3 and the subsequent Figures 5-10 for different experiments.

Table 3. AUC comparison of different deep learning configurations.

#	Experiment	Dataset				
		LHMC	UCM	NLST	Kaggle 1	Kaggle 2
1.	Our model (described in Table 2)	0.8728	0.8262	0.8756	0.8235	0.8394
2.	Our model with only one nodule as input	0.8330	0.8078	0.8562	0.8062	0.8271
3.	Our model with larger nodule cubes (64x64x64)	0.8534	0.7860	0.8657	0.8159	0.8381
4.	Our model with smaller dropout (0.6)	0.8769	0.8149	0.8754	0.8207	0.8177
5.	Our model with only image input	0.8512	0.7861	0.8563	0.7632	0.8207
6.	Our model with only numeric inputs	0.8237	0.7633	0.7659	0.7757	0.8016
7.	DenseNet	0.7966	0.7266	0.8052	0.7317	0.7855

Experiment: Single nodule per volume
In this experiment, we keep the same configuration as in Table 2 with the difference being that we use a single (largest nodule) instead of 10 as input. The obtained performances are shown in Figure 5 and in rows #1 and #2 in Table 3.

Experiment: Larger patch around a nodule
We demonstrate that taking more context (larger cubes around nodule locations) does not improve performance (see Figure 6; and rows #1 and #3 in Table 3). This can be explained by the fact that taking a larger cube around a nodule increases the number of parameters and also includes additional non-relevant context.
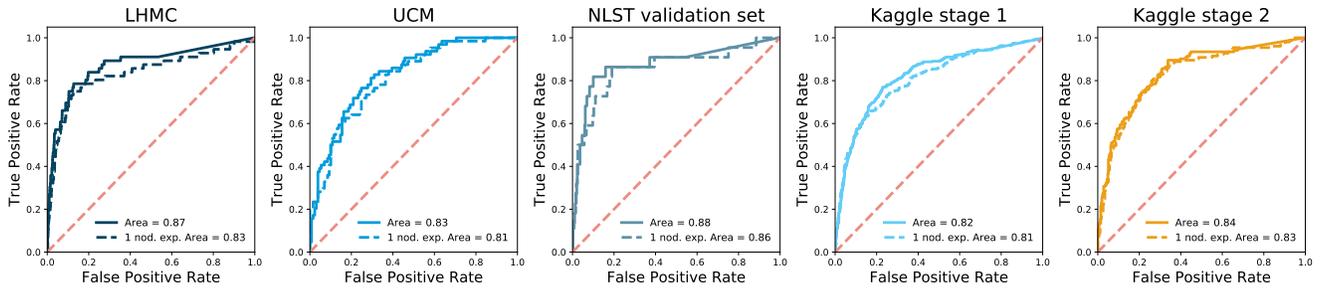
Figure 5. Performance comparison of the best model as described in Table 2 that uses ten nodules as compared to the model of using only one nodule.
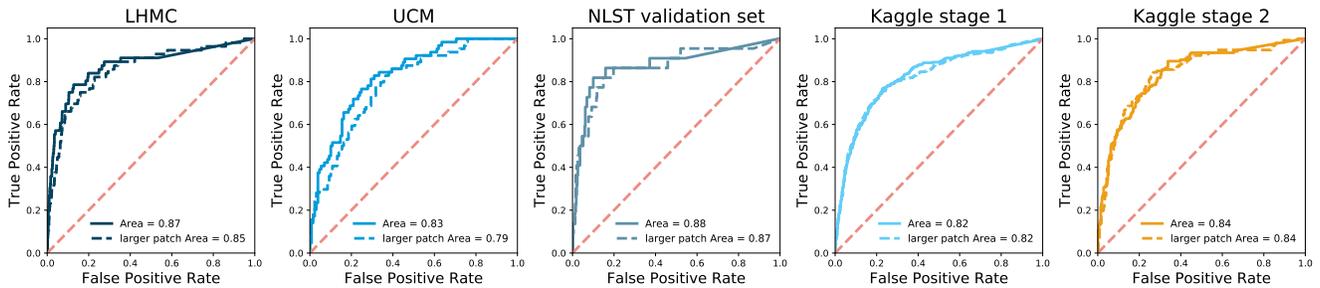

Figure 6. Performance comparison of the best model as described in Table 2 that uses nodule cubes of size 32x32x32, followed by random augmentation crop 28x28x28 and an experiment with nodule cubes of size 64x64x64, followed by random augmentation crop 55x55x55.

Experiment: Smaller dropout rate

In this experiment, we keep the same configuration as in Table 2 with the difference being that we have a smaller dropout of 0.6 rather than one from $\{0.7, 0.8, 0.9\}$. The comparison results are shown in Figure 7 and in rows #1 and #4 in Table 3. The experiment demonstrates the having high dropout helps in achieving better performance. It is also worth mentioning that adding more nodules slightly reduces the role of the dropout as having more nodules plays a role of a "regularizer".
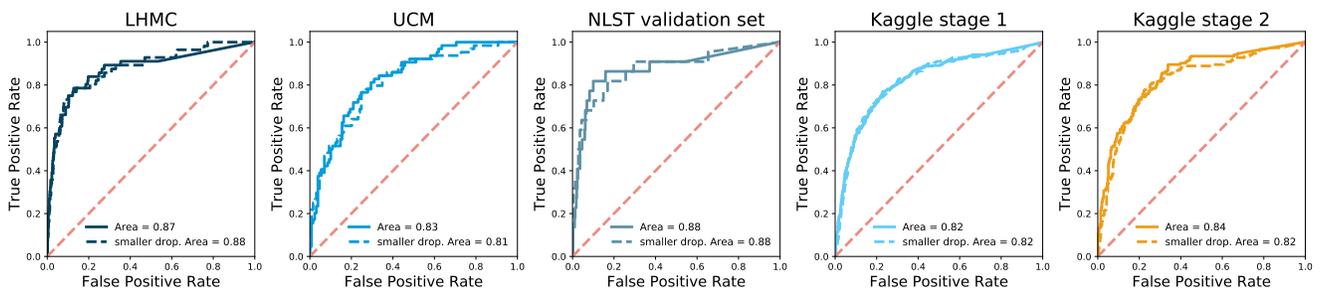

Figure 7. Performance comparison of the best model as described in Table 2 that uses dropout of at least 0.7 and an experiment with dropout 0.6.

Experiment: Only image input

In this experiment, we keep the same configuration as in Table 2 with the difference being that only the image part is used as an input in the neural network (rows #1 and #5 in Table 3). We don't use the numeric nodule description inputs: nodule radiuses, svm scores (confidence level of detected nodule as provided by the SVM algorithm used by the nodule detector), x, y and z coordinates obtained from nodule detection stage of the algorithm. Figure 8 shows that the optimal is obtained by combining the image input combined with the numeric nodule descriptor. We should stress that the nodule descriptors are obtained automatically by the nodule detector without any human intervention.
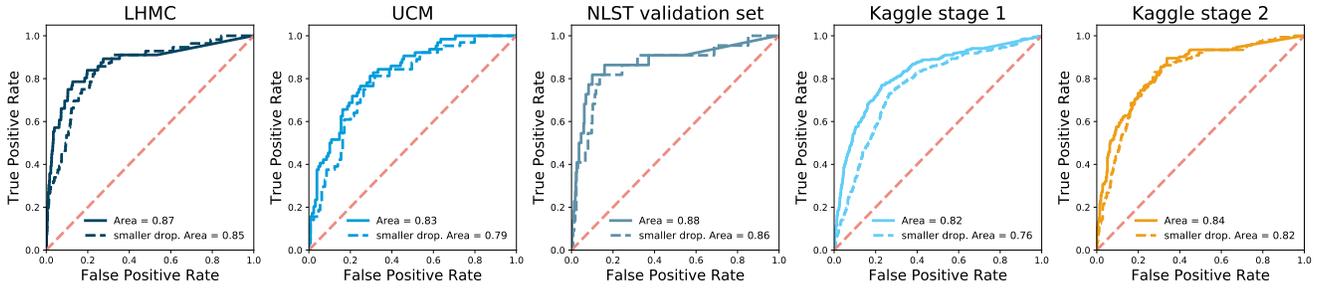
Figure 8. Performance comparison of the best model as described in Table 2 and an experiment without the numeric nodule descriptors.

Experiment: Only numeric inputs

In order to examine the effect of the numeric nodule descriptors without the image part, we conduct an experiment where only this information is used. More precisely, we construct ten single layer neural networks, one for each nodule, using the nodule descriptors as input. These single layer networks have a sigmoid output that generates the risk score for each nodule. Consequently, using a Global Max Pooling layer, we produce the malignancy score for the whole scan, which is effectively the largest nodule risk score. This network is trained in a similar manner as our best model, using the verified cancer cases as labels. Although, this experiment achieves a meaningful performance, it performs worse than the combination of image and numeric nodule description data (rows #1 and #6 in Table 3 and Figure 9).
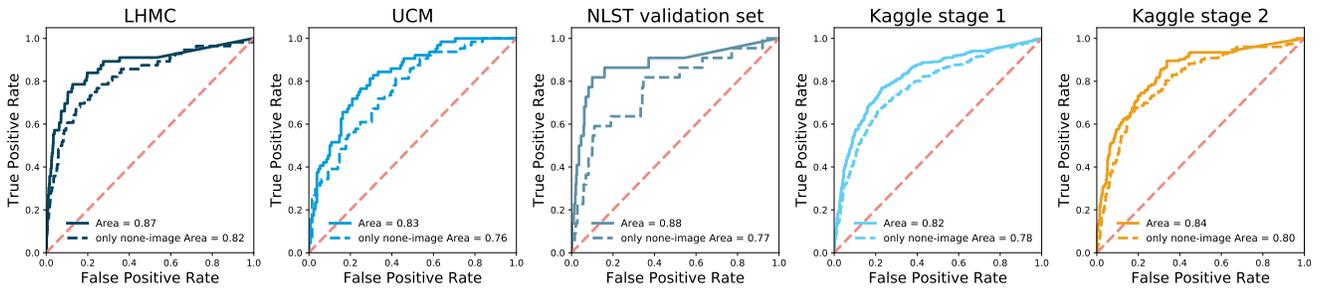

Figure 9. Performance comparison of the best model as described in Table 2 and an experiment without the image part (only numeric nodule descriptors as inputs).

Experiment: DenseNets

Finally, we compare the best model configuration as described in Table 2, which relies on ResNet architecture, with a model that uses the DenseNet architecture[43]. Our experiments (Figure 10 and rows #1 and #7 in Table 3) demonstrate that DenseNet performs worse than our ResNet-like architecture.
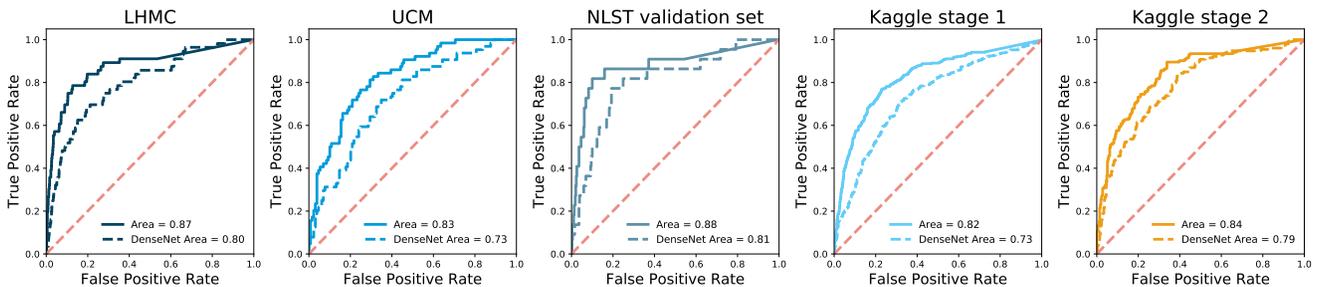

Figure 10. Performance comparison of the best model as described in Table 2 and an experiment with DenseNet architecture[43].