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

Stojan Trajanovski[1]

Philips Research, 5656 AE Eindhoven, The Netherlands.

Dimitrios Mavroeidis[1]

Philips Research, 5656 AE Eindhoven, The Netherlands.

Christine Leon Swisher[1]

Human Longevity, Inc., San Diego, CA 92121, USA. Work done while with Philips Research North America, Cambridge, MA 02141, USA.

Binyam Gebrekidan Gebre

Philips Research, 5656 AE Eindhoven, The Netherlands.

Bastiaan S. Veeling

Machine Learning lab, University of Amsterdam, 1090 GH Amsterdam and Philips Research, 5656 AE Eindhoven, The Netherlands.

Rafael Wiemker

Philips Research, 22335 Hamburg, Germany.

Tobias Klinder

Philips Research, 22335 Hamburg, Germany.

Amir Tahmasebi

Philips Research North America, Cambridge, MA 02141, USA.

Shawn M. Regis

Lahey Hospital & Medical Center, Burlington, MA 01805, USA.

Christoph Wald

Lahey Hospital & Medical Center, Burlington, MA 01805, USA.

[1]The first three authors have contributed equally.

Brady J. McKee

Lahey Hospital & Medical Center, Burlington, MA 01805, USA.

Sebastian Flacke

Lahey Hospital & Medical Center, Burlington, MA 01805, USA.

Heber MacMahon

Department of Radiology, University of Chicago, Chicago, IL 60637, USA.

Homer Pien

Philips Research North America, Cambridge, MA 02141, USA.

Abstract

Purpose: Lung cancer is the leading cause of cancer mortality in the US, responsible for more deaths than breast, prostate, colon and pancreas cancer combined and large population studies have indicated that low-dose computed tomography (CT) screening of the chest can significantly reduce this death rate. Recently, the usefulness of Deep Learning (DL) models for lung cancer risk assessment has been demonstrated. However, in many cases model performances are evaluated on small/medium size test sets, thus not providing strong model generalization and stability guarantees which are necessary for clinical adoption. In this work, our goal is to contribute towards clinical adoption by investigating a deep learning framework on larger and heterogeneous datasets while also comparing to state-of-the-art models.

Methods: Three low-dose CT lung cancer screening datasets were used: National Lung Screening Trial (NLST, n=3410), Lahey Hospital and Medical Center (LHMC, n=3154) data, Kaggle competition data (from both stages, n=1397+505) and the University of Chicago data (UCM, a subset of NLST, annotated by radiologists, n=132). At the first stage, our framework employs a nodule detector; while in the second stage, we use both the image context around the nodules and nodule features as inputs to a neural network that estimates the malignancy risk for the entire CT scan. We trained our algorithm on a part of the NLST dataset, and validated it on the other datasets. Special care was taken to ensure there was no patient overlap between the

train and validation sets.

Results and Conclusions: The proposed deep learning model is shown to: (a) generalize well across all three data sets, achieving AUC between 86% to 94%, with our external test-set (LHMC) being at least twice as large compared to other works; (b) have better performance than the widely accepted PanCan Risk Model, achieving 6 and 9% better AUC score in our two test sets; (c) have improved performance compared to the state-of-the-art represented by the winners of the Kaggle Data Science Bowl 2017 competition on lung cancer screening; (d) have comparable performance to radiologists in estimating cancer risk at a patient level.

---

1. Introduction

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]. Average five year survival for lung cancer 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. However, early stage lung cancer (stage I) has a five-year survival of 60-75%. In 2011, the National Lung Screening Trial (NLST) demonstrated that lung cancer mortality can be reduced by at least 20% using an annual screening program of high-risk populations with low-dose computed tomography (CT) of the chest [3]. A even greater mortality reduction has been recently reported in the European Nelson trial [4].

The process of lung cancer risk assessment typically involves two steps: nodule detection and malignancy risk assessment. Lung nodules are small tissue masses that are located in the lungs and since they are not uncommon, a second step is required to evaluate the cancer malignancy risk based on the identified nodules. Cancer malignancy assessment is commonly based on observed changes in the nodule characteristics (like growth in size) between scans that are taken in regular time intervals. In this work we focus on single-scan criteria, like the ones used in the *PanCan risk model* [30] that consider several factors such as the nodule location, size and shape (for further details one can look into the Lung-RADS$^{\text{TM}}$ [32] protocol). In case a suspicious nodule is identified in the screening process, further evaluation steps are taken (like biopsy) to verify malignancy.

A complete implementation of lung cancer screening (LCS) programs at a national level could result in a large volume of CT lung screening (CTLS)

scans that need to be assessed by radiologists, e.g., the insurance mandated high-risk criteria are met by millions of Americans [5]. This highlights the potential utility of image analysis tools that can help radiologists to assess malignancy risk associated with a CTLS scan and make recommendations for further work-up, e.g. pulmonologic, oncologic or surgical evaluation.

There exists substantial literature related to both nodule detection and cancer malignancy estimation (see e.g., [13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 15, 25, 26, 27, 28, 29], with recent works achieving significant performance improvements using DL and Convolutional Neural Network (CNN) architectures. Moreover, a recent data science competition (Kaggle, Data Science Bowl 2017 [34]) was organized on the topic of malignancy risk assessment that attracted a significant attention from the research community. The authors of the winning entry Liao *et al.* [35] proposed a 3D CNN network for nodule detection, using LUng Noduel Analysis 2016 (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. Ardila *et al.* [37] proposed a cancer malignancy estimation framework with two main differences compared to related work: they do not rely on only nodule locations since they additionally use a 3D neural network to extract features from the full scan and their model can use a previous scan of a patient when available to determine the cancer malignancy probability.

The related works on cancer malignancy assessment provide indications that machine learning (deep learning) models can achieve good performances for tasks related to lung cancer risk assessment. However, these works are either based on small datasets or do not compare with other state-of-the-art methods (for example [37] where no comparisons are presented to any prior machine learning works for cancer malignancy estimation). Larger scale experiments and better comparison to state-of-the-art models are however essential to provide further evidence for clinical adoption.

In this paper, we propose a two-stage Machine Learning framework that estimates the cancer risk associated with a given CTLS scan. The first relies on nodule detection [36, 35] that identifies nodules contained within a scan. The second stage employs a neural network inspired by the ResNet architecture [39] and regularized using dropout [40] that performs the cancer risk assessment of the whole CTLS scan. Our framework is evaluated against three criteria (i) robustness: we show that our framework achieves consistent performance across different low-dose CT datasets, (ii) performance against

state-of-the-art: we show that our framework has improved performance over existing methods, (iii) performance compared to radiologists: we show that our model has comparable performance to a panel of six radiologists and (iv) performance with several parameter choices: we experimentally illustrate that nodule detection and malignancy assessment can be two independent processes and promote the re-use of off-the-shelve nodule detectors or existing products as a first step for cancer malignancy assessment. To the best of our knowledge, no other study has utilized all these benchmarks simultaneously.

2. Methods

2.1. Data Sets

In order to evaluate the performance of our framework, we use CTLS datasets from several sources: NLST [3], LHMC, Kaggle [34] (from both competition stages) and University of Chicago (UCM) data (NLST subset with radiologist annotations). The main data characteristics are summarized in Table 1. The CTLS datasets we have used in our analysis come from heterogeneous sources (different hospitals, image quality, reconstruction filters, etc.) and enabled the validation of the generalization capacity of our framework. In order to facilitate model training, we have used all the diagnosed cancer CTLS scans from the NLST dataset and a subset of the benign cases. This is a common practice in highly imbalanced class-distributions to undersample the majority class or oversample the majority class. It should be noted that when validating our model we ensured that there was no patient overlap between the train and validation sets, for example by removing the (cancer and non-cancer) patients that are contained in the UCM dataset.

Table 1: Data used in our analysis.

	Number of volumes			Metadata		
Dataset	Total	positive	train or valid.	nodule annotations	Lung-RADS^{TM} classification	radiologists' scores
NLST	3410	680	train our model	yes	no	no
LHMC	3154	43	valid	yes	yes	no
UCM	132	28	valid	yes	no	yes (for 81 volumes)
Kaggle (stage 1) train	1397	362	train model [35]	no	no	no
Kaggle (stage 2)	505	153	valid	no	no	no

We trained our model on the NLST data [3] (3410 volumes, containing 680 hundred biopsy-diagnosed cancer cases) since this dataset has the largest number of cancer cases. The NLST-trained model was subsequently validated with the other datasets. When validating model performance for the UCM

Table 2: Nodule and Patient characteristics used in the PanCan model comparison

	Malignant Nodules		Benign Nodules	
	UCM	LHMC	UCM	LHMC
total number of nodules evaluated	28	14	104	627
solid	25	7	93	480
part-solid	3	7	5	92
ground-glass	0	0	6	55
spiculated	20	3	20	15
upper lobe	19	8	40	201
average diameter	13.85mm±5.11	14.14mm±7.95	9.56mm±4.37	6.75mm±4.02
average age	63.82±5.81	63.35±6.79	62.65±5.34	64.32±5.81
num. female,num. male	12,16	7,7	33,71	164,239
family history	7	3	22	53
average num. nodules	1.5	2.42	1.84	1.59
num. emphysema	15	9	45	263

dataset, we always excluded from the NLST training data the cancer and non-cancer patients that were included in the UCM study. The lung cancer screening dataset provided by LHMC contains 3154 CTLS patient scans (with 43 biopsy confirmed 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 the LHMC dataset, but the detailed nodule information allows us to compare our framework with other models from the literature that rely on such nodule-level information [30, 32]. Furthermore, UCM has provided additional annotations for 132 volumes of the NLST data (that contain 28 cancer cases), that allow us to compare our model with radiologists' assessment as well as the *PanCan risk model*. Finally, we use the data from a recent lung cancer competition (National Data Science Bowl 2017) organized and hosted by Kaggle [34]. In the first stage of the competition, 1397 CTLS volumes were provided for training data (with 362 diagnosed cancer cases) and for validation 198 CTLS volumes (with 57 diagnosed cancer cases), which are used to train Liao *et al.* model [35] (both nodule detection and cancer malignancy estimation); while in the second stage 505 volumes were provided (with 153 cancer cases). In all our datasets, cancer cases were confirmed with diagnostic tests (like biopsy), so it is almost certain that the labeling is unambiguous, however, for the non-cancer cases there is a possibility that a patient left the study and developed cancer later on.

Table 2 summarizes the nodule and patient characteristics for the data used in the *PanCan risk model* comparison. For the UCM dataset we report

the distribution for one nodule per scan, the one for which the *PanCan risk model* produces the highest cancer risk probability. For the LHMC dataset we report the characteristics for all nodules contained in the 417 scans used in the comparison. Moreover, since the annotations (diagnosed cancer cases) are available only at the scan level, we report the malignant nodule distribution based on the nodule that produces the maximum cancer malignancy probability for the respective scan.

The estimated CTDI vol. for the NLST dataset was 2.9mGy [45], while for the LHMC it was 1.37mGy (computed over a subset of 514 LHMC scans for which this information was available). The NLST dataset contains both soft-tissue and sharp/medium reconstruction kernels while in the LHMC dataset, the majority of the reconstruction filters was soft-tissue 2818/3154. For the UCM dataaset the number of soft-tissue reconstruction cases was 79/132. For the Kaggle dataset, there is no information available about scan dose or reconstruction filters used.

Data overlap safeguard.. In order to evaluate and compare machine learning models, we need to investigate whether there is an overlap between training and test datasets. This becomes especially important if we want to compare the performance of our framework to the state-of-the-art Liao *et al.* model, since we need to ensure that there is no overlap between the data used to train this model and our test sets. The main challenge is that the Liao *et al.* model was trained using Kaggle competition data, a dataset that specifies as data sponsors both the NLST and LHMC. For this reason, we checked for any possible overlaps between the data that was used to train our model (NLST) and Liao et al. model (Kaggle stage 1) with the test sets where we report performances in this paper (UCM, LHMC and Kaggle stage 2).

The data overlaps that are important for our work are:

- Kaggle stage 1 LHMC: since Liao *et al.* model was trained on Kaggle stage 1 and LHMC was a data sponsor
- Kaggle stage 1 UCM: since Liao *et al.* model was trained on Kaggle stage 1 and NLST was a data sponsor
- NLST - Kaggle stage 2: since our model was trained on NLST data and NLST was a data sponsor for Kaggle

To resolve the potential data overlap issue, we compared the CT images (pixel data in the dicom files) between the aforementioned dataset pairs.

The CT images that were found to have equal voxel values, were subsequently removed from the test sets (UCM, LHMC and Kaggle stage 2). This comparison was performed at the patient level and all scans from the NLST patients used to train our model were taken into account, as well as all the scans provided by LHMC. When an overlap is detected, all scans from the respective patient were removed from the test sets. In order to also consider possible ad-hoc voxel value perturbations or corruption, we employed the results of the nodule detector and compared two scans when they had at least one nodule with a similar (x, y, z) location and size. In these cases, the two scans were checked for voxel value (CT image) similarity but also visually verified.

By using this analysis we were able to identify significant overlaps mostly between the dataset pairs (Kaggle stage 1 - NLST) and (Kaggle stage 2 - LHMC). These overlaps do not affect our analysis, since we do not report NLST or Kaggle stage 1 performances for any models and similarly no model is trained on Kaggle stage 2 or LHMC data. The overlaps identified in the relevant dataset pairs (Kaggle stage 1 - LHMC), (Kaggle stage 1 - UCM), (NLST - Kaggle stage 2) were removed from the test sets used in this paper (UCM, LHMC and Kaggle stage 2). The data set sizes reported in Table 1 are after overlap removal.

Developments in AI research rely crucially on large scale comparisons of state-of-the-art models that guide further research works. Thus, we consider our analysis to be a useful contribution to the research literature because it allows us to compare with state of the art models developed for the Kaggle competition. Moreover, we expect similar challenges to be present for Lung Cancer Screening models that are trained using the NLST dataset since this is a multi-center dataset with data contributions from several hospitals. For example, in the the paper of Ardila *et al.* [37] the Lung Cancer Screening model was trained on the NLST dataset [3] and validated using data from the from the Northwestern Medicine, while Northwestern University is a data sponsor of the NLST dataset [3] (Supplementary Material; protocol file, page 3). The issue of the potential train/test data overlap was not discussed in their work.

2.2. Machine Learning Framework for Cancer Risk Assessment

Model architecture. In this work, we propose a two-stage machine learning framework for cancer risk assessment that follows the two stages that a radiologist would take for assessing a scan. In the first stage, we employ a nodule

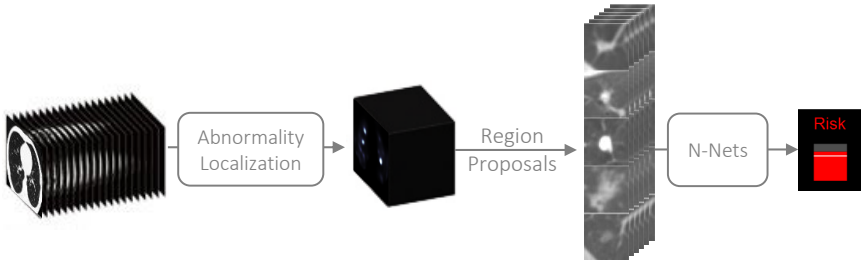
detector to identify the nodules that are contained in a CTLS scan, while in the second stage we use the ten largest nodules identified by the nodule detector as input to a deep and wide neural network that assesses their cancer risk. The decision to use the ten largest nodules was based on the optimal 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 1.

We evaluate this framework with two different nodule detectors, one based on hierarchical Support Vector Machines (SVMs) by Bergtholdt *et al.* [36] (further referred to as *detect_SVM*) and another based on deep neural network semantic segmentation by Liao *et al.* [35] (further referred to as *detect_CNN*). Both are based on LUNA16 dataset [12] of confirmed nodule cases, while the later contains additional annotated cases of very big nodules (or already developed tumors) from the Kaggle dataset. More details about the nodule detectors are given in the above mentioned manuscripts [35, 36].

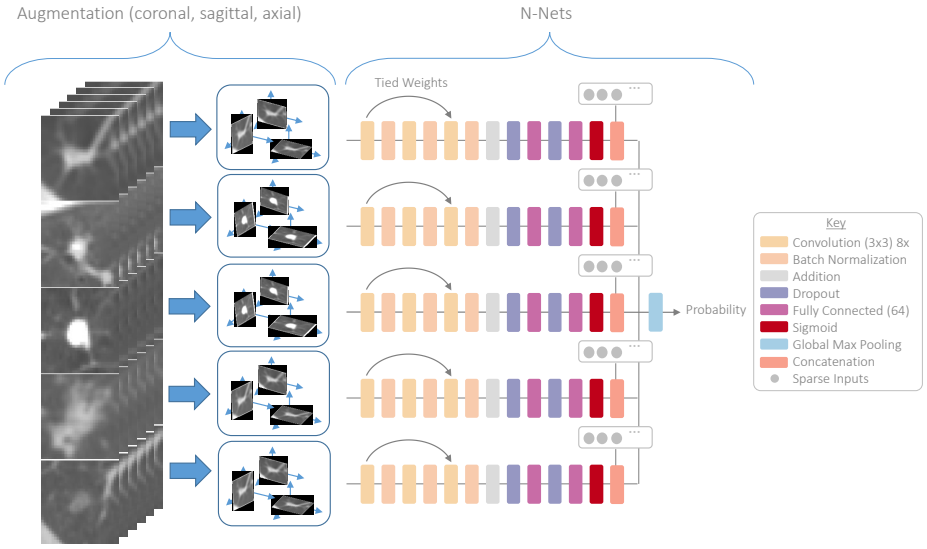
The nodule detector provides us with the nodule locations in all three dimensions: x, y and z as well as additional information such as the nodule size (e.g., radius in mm), and the confidence of the suggestion - given by the nodule detector. We refer to these parameters as nodule metadata.

Based on the output from the previous stage, we extract from the CTLS scan localized cubes of 32x32x32mm^3 around a nodule (and since we employ isotropic resampling to 1mm^3 each voxel corresponds to 1mm^3). This gives us sufficient context for the experiments as we found that smaller or larger cubes do not improve and can even degrade performance. One way to interpret this observation is that for nodules larger than 32mm the cancer risk is dominated by the radius and position of the nodule that are provided as additional input to the network along with the image data. Additionally, during training, a random subimage of 28x28x28mm^3 out of the extracted 32x32x32mm^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 28x28x28mm^3 cube we extract three different 2D projections, as channels, namely coronal, sagittal, and transverse, thus ending up with 3 times 28x28 input per nodule for the neural network (Figure 1).

Moreover, for each nodule, we use additional features, such as nodule radius, and confidence score (confidence level of a detected nodule as provided by the algorithm used for nodule detection) 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. In



A



B

Figure 1: 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. B: The network architecture that is applied to the cube around the detected nodules. For simplicity we depict here only 5 nodules instead of 10 that are actually employed by our model

the experiments that use the *detect_SVM* nodule detector we employ one additional feature, nodule sphericity. The reason for this difference is that the nodule-sphericity feature is not provided by the *detect_CNN* nodule detector. Different volumes have different number of nodules. In the experiments, we

Table 3: 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), stride 1	1.
3. Conv Layer + BN	(3x3, 8x), stride 1	2.
4. Conv Layer + BN	(3x3, 8x), stride 1	3.
5. Conv Layer + BN	(3x3, 8x), stride 1	1.
6. Addition/Merge + BN	-	4., 5.
7. Dropout + BN		6.
8. Dense + BN	(64)	7.
9. Dropout + BN		8.
10. Dense + BN	(64)	9.
11. Numeric input	(10x1)	- Radius
12. Numeric input	(30x1)	- x, y, z nodule coordinates
13. Numeric input	(10x1)	- confidence score
14. Addition/Merge	-	10., 11., 12., 13.
15. Dense + sigmoid	(1)	14.
16. GlobalMaxPool	(10)	15.

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. We use a ResNet-like [39] 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 1. More details of the exact layer configuration of the neural network are given in Table 3. We used 3x3 kernels for convolutional neural network blocks with 8 channels with stride 1, intertwined with batch normalization and additional connections for realizing the ResNet-blocks (see inputs 5. and 6. in Table 3), 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. Our network

architecture has 407 766 parameters. We have employed the aforementioned architecture with both nodule detectors, with the only difference being the dropout rate. More precisely, when using the *detect_SVM* nodule detector we set the dropout rate to very high values (0.7-0.9) that were necessary to obtain optimal performances, while for the *detect_CNN* nodule detector the dropout rate was set to a much smaller value (0.25).

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. With our approach, model training is feasible simply by using patient diagnosis information and does not require knowledge about which lung nodule was malignant.

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 evaluated the performance of our model with respect to verified cancer diagnosis at the volume level. The model that employs the *detect_CNN* nodule detector, was trained using 5-fold cross-validation on the NLST data excluding all patients used in the UCM dataset and during inference the average prediction of the 5-folds was used. For the model that employs the *detect_SVM* nodule detector, we have employed again the NLST data for training and we have used three model versions, one trained using 90% of the NLST training data and 0.9 dropout, one using the full NLST data excluding the full UCM patient list (132 patients) and 0.8 dropout and the full NLST data excluding the UCM subset rated by Radiologists (81 patients) and 0.7 dropout. This was done because we observed that the SVM-nodule model had reduced performance when using smaller NLST subsets for training. Thus, when testing the model performance, we employed the larger version possible from NSLT, i.e. we used the NLST dataset without the UCM patients only when testing for that data subset. This behavior was not observed when we employed the DL-nodule model and we were able to use a single model excluding all UCM patients from NLST without observing reduced performance in our test sets. Moreover, we should note that the convergence of our models when using high dropout rate (even 0.9) was consistent, possibly due to the fact that we used the nodule metadata at the penultimate layer of the neural network

architecture. The high dropout rate for *detect_SVM* can also be attributed to the fact that*detect_SVM* tends to detect smaller nodules compared to *detect_CNN* that may be less relevant for the lung cancer risk assessment process. We used Adam optimizer [42] with a learning rate of $1e-3$. For the N-Net with *detect_CNN*, training was stopped when the average loss over the 5-folds was not further improving. This means that the same epoch was used for all 5-folds which was the 730th. Similar approach was used for the N-net models that employs the nodule detector from *detect_SVM*, with the number of epochs being 600 for the model trained with the full NLST data, 610 epochs for the model trained excluding the UCM patient subset contained in the radiologist evaluation set and 640 epochs for the model that excluded all the UCM patients.

We explored the individual contribution of the architecture's attributes with various ablation experiments. More precisely, 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 [43] and DenseNets [44]. The results suggested that there is no benefit in these architectures and the proposed one (in Table 3 and Figure 1) performs better than the alternative architectures or hyper-parameters. Similar hyper-parameter exploration was conducted for both nodule detectors, resulting in the same neural network architecture with the only difference being the optimal dropout rate, which was lower (0.25 for the *detect_CNN* nodule detector vs. 0.7-0.9 for the *detect_SVM* nodule detector).

2.3. PanCan Risk Model

To empirically validate our framework, we employ a model developed at the Vancouver General Hospital for nodule malignancy estimation [30], which is mentioned in the Lung-RADS$^{\text{TM}}$ [32] guidelines as a recommended tool for assessing nodule malignancy risk. This method uses a single patient 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, 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: number of nodules in the CTLS scan;

and four nodule specific image-based features: size of a nodule (diameter) - which is longest in-slice axis, type of the nodule (nonsolid, part-solid, or solid), location of the nodule in the upper lobe (true or false), and nodule spiculation (true or false).

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

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 model* (rather than taking the mean, minimum scores etc. of a nodule per study).

2.4. Radiologists Predictions

To compare our results to radiologist performance, an observer study was conducted at UCM using 81 out of the 132 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 10 malignant and 71 benign cases. Each selected case had to have at least one nodule within the range of 6-25mm. Besides nodule size distribution matching, the selection covered nodule types of all categories except for calcified nodules. Three senior (radiologists 1, 2 and 3) and three junior (radiologists 4, 5 and 6) 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. This information included 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.

3. Results

3.1. Performance robustness and comparison with the state-of-the-art

In order to assess the robustness of our model we have evaluated its performance across our different test-sets. In Figures 2A, 2B, 2C, 2D we can observe that the performance of our framework was stable across the different datasets and achieved an AUC (Area Under the Curve) score between 86%-94%. 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. This is an important point since robustness and generalization are essential for clinical adoption. The lower performance on the UCM datasets can be attributed to the fact that the cases included in this study are all characterized as: "Positive, Change Unspecified, nodule(s) >= 4 mm or enlarging nodule(s), mass(es), other non-specific abnormalities suspicious for lung cancer" according to the "Result of isolation screen" information for the respective scan in the NLST study. This means that the scans include one or more nodules that require assessment and monitoring. This is different form a standard screening population that will include a large number of easy to assess negative screening cases that would improve the performance of a cancer risk assessment model. Our evaluation is more extensive than the majority of related works that commonly use smaller and less diverse datasets. The model is trained on NLST dataset and it does not require additional re-training and is just evaluated on the remaining data sets.

Figure 2E illustrates that the performance of our framework is better than the winner's of the Kaggle Data Science Bowl challenge on Lung Cancer Screening (Liao *et al.* model), with a performance difference of 1.2% and associated p-value $P = .0448$ (computed using a two-sided permutation test [41]). It should be noted that the Liao *et al.* model has also robust performance across our different test-sets and the performance improvement of our model becomes apparent only when we aggregate all our test sets together. The performance robustness was also confirmed by an independent validation at Moscow Radiology Center, where our framework achieved an AUC of 93% [46].

3.2. The influence of the choice of the nodule detection

In Figure 2, it can be observed that although the same neural network architecture (referred to as N-Net in the figures) was used with both nod-

ule detectors of *detect_CNN* [35] and *detect_SVM* [36], there is a substantial performance difference. In order to understand this behavior, we looked at cases with different results and we have observed that the *detect_SVM* nodule detector missed the malignant nodule in some cases, especially where the nodule was already a large developed tumor. Such large tumors are not covered by LIDC data which was used to train the *detect_SVM* nodule detector, while they were part of the *detect_CNN* nodule detector that has used additional annotations from Kaggle stage 1 data. In fact the authors of *detect_CNN* have highlighted the lack of large tumors as a potential limitation of the LIDC dataset for training effective nodule detectors for Lung Cancer Screening. These cases of missed malignant nodules, although small in number, significantly affected the ROC curves since these cancer cases were associated with a very low probability. In order to make this argument more quantitative, we also report here the performances of N-NET with *detect_SVM* when we remove the cases with malignant nodules that were missed by *detect_SVM* and identified by the nodule detector proposed in *detect_CNN*. In summary the AUC score for LHMC becomes 91.6% (two cases with malignant nodules that were identified by detect_CNN and not *detect_SVM* removed from the analysis), Kaggle stage 2 becomes 86% (five such cases removed), UCM becomes 82.4% (two such cases removed).

In Figure 5 we provide some examples where [35] nodule detector identifies the malignant nodule that is missed by [36], highlighting the need for nodule datasets (like the LIDC data) to include larger malignant nodules.

3.3. Comparison with radiologists performance

Figure 3C 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 81 volumes (different patients) of the UCM data out of which 10 correspond to verified cancer cases. Our algorithm shows a comparable and often better performance than the performance of the radiologists. The predictions of one radiologist that has slightly better AUC scores, namely Radiologist 2 in Figure 3C, cannot be considered statistically significant, since the p-values, when compared to our methods, are much higher from what is normally considered statistically significant. It can also be observed that if we set the sensitivity threshold to 100% then only one radiologist, namely Radiologist 5 in Figure 3C, is able to achieve better specificity compared to our model.

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 LCS (i.e. no cancer cases are missed). It should be noted that our work is one of the few studies [28] in the literature where such a comparison to radiologists is performed.

3.4. Comparison with the PanCan Risk Model

The results, presented in Figures 3A and 3B show that our proposed model outperforms the *PanCan risk model* [30] by approximately 9% and 6% AUC in UCM and LHMC datasets, respectively. The performance is clearly better along all points of the RO curve but p-values cannot prove statistical significance since the number of samples is small (order of hundreds).

In order to provide a better understanding of the model performances for various Lung-RADS^{TM} categories, we conduct two experiments for fixed sensitivity and specificity thresholds. In the first experiment, in Figure 3E, we fix the sensitivity to 93%, i.e., requiring that the models identify in a single scan 93% of all malignant nodules and then compare their specificity for the different Lung-RADS^{TM} categories. At this level of sensitivity our model achieves a specificity of 81% vs. specificity of 69% for the *PanCan risk model*. In Figure 3D, we fix the specificity to 80% and observe that our model achieves a sensitivity of 93% vs. 79% for the *PanCan risk model*.

In Figure 4 we provide some specific examples where our DNN model ranks the nodule risk score more accurately than the *PanCan risk model*. We hypothesize that the filters of our CNN model can capture in a better way the shape-characteristics of a malignant nodule beyond a simple scalar value that is used in the *PanCan risk model* to account for the different nodule characteristics (like spiculation, etc.).

4. Discussion

In recent years, several research papers have proposed solutions for the problems of nodule detection and nodule malignancy assessment. However, the evaluation of these models is done usually in datasets of much smaller scale that what is used in this paper. For example van Riel et al. [28] uses a dataset of 300 CT scans to train and evaluate their model while in this study we employ 20x more scans from a combination of public and private data sources. This allows us to have more confidence about the robustness and generalization capacity of our framework, which was also

confirmed by an independent validation at Moscow Radiology Center, where our framework achieved an AUC of 93% [46]. The only exception to this rule is the paper from Ardila *et al.* [37], where the dataset sizes are comparable to our work, although still the clinical dataset we have used (LHMC) is almost twice as large. The disadvantage of Ardila *et al.* [37] however is that they did not benchmark their method against state-of-the-art Machine Learning methods and what we observe in this work is that both our model and also the DSB Kaggle competition winners (Liao *et al.* [35]) model achieve comparable results with much simpler models.

In this work, we evaluated the performance of the DSB Kaggle competition winners (Liao *et al.* [35]) in all our datasets, thus allowing us to make a thorough comparison of our model performance against state-of-the-art approaches. Interestingly the Liao *et al.* model had strong performances in most datasets but our model can achieve a statistically significant performance improvement on the aggregated test-set. Beyond the performance comparison, validating the Liao *et al.* model on additional test sets serves as further evidence regarding the robustness of modern deep learning approaches and also the usefulness of data challenges and competitions to advance scientific research.

From the methodological point of view, our work contributes to a better understanding of the type of information that is needed to train highly performant Deep Learning models for cancer malignancy estimation. The two main paradigms are (i) models that employ solely nodule location information and patient outcomes (diagnosed-cancer or no-cancer) as information to train the model, (ii) models that employ additional information about nodule characteristics (such as level of spiculation, lobulation, etc.). We employed only the nodule locations and patient diagnostic outcomes to train our model. To the extent of our knowledge the only other works that rely on the same information is the recent Ardila *et al.* model and the Liao *et al.* model, i.e. the Kaggle DSB winners model that scored better against approaches that employed additional nodule characteristics. The results presented in this paper illustrate that nodule locations and patient diagnostic outcomes suffices to build high-performance deep learning models for lung cancer malignancy estimation.

Another interesting methodological question is related to whether nodule detection and malignancy estimation should be trained in a common end-to-end model or they can be two separate tasks. The models proposed by Ardila *et al.* and Liao *et al.* are trained in an end-to-end manner, while

in our work we have used a two-stage model where the training of the second stage is separate from the nodule detection. Interestingly, our results demonstrate that a two-stage approach can be very successful and both nodule detectors performed very well on datasets that they have not seen before (like NLST and LHMC). Moreover, a two-stage approach can offer greater transparency and explainability of the system's decision logic. Our results can lead to follow-ups, since patient diagnosis results can be retrieved from patient records and do not have the level of ambiguity that is sometimes associated with radiologist reports. Moreover, the availability of a substantial number of nodule detectors as a result of the LUNA16 challenge means that a clinical research center that performs lung cancer screening has the potential to build and validate their own lung cancer screening model based on the principles outlined in this work.

5. Conclusion

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 that 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 is shown to have (i) robust performance across three low-dose CT dataset, (ii) improved performance compared to state-of-the-art models and (iii) comparable performance to a panel of six radiologists. 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] R. Siegel, J. Ma, Z. Zou, and A. Jemal, "Cancer statistics, 2014", CA: A Cancer Journal for Clinicians **64**, 9 (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 England Journal of Medicine **365**, 395 (2011) .

[4] International Association for the study of Lung Cancer (IASLC), "IASLC 19th World Conference on Lung Cancer (Collection of abstracts)," https://wclc2018.iaslc.org/ (2018) .

[5] H.J. de Koning, R. Meza, S.K. Plevritis, K. ten Haaf 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 of Internal Medicine **160**, 311 (2014).

[6] F. Preteux, "A non-stationary markovian modeling for the lung nodule detection in ct", in *Computer Assisted Radiology / Computergestützte Radiologie: CAR '91 Computer Assisted Radiology*, edited by H. U. Lemke, M. L. Rhodes, C. C. Jaffe, and R. Felix (Springer Berlin Heidelberg, Berlin, Heidelberg, 1991) pp. 199–204.

[7] S. Benedict Lo, J. Lin, M. Freedman, and S. Ki Mun, "Computer-assisted diagnosis of lung nodule detection using artificial convoultion neural network", in *Proc. SPIE*, (1993) .

[8] T. Messay, R. C. Hardie, and S. K. Rogers, "A new computationally efficient CAD system for pulmonary nodule detection in CT imagery", Medical Image Analysis **14**, 390 (2010).

[9] N. Camarlinghi, I. Gori, A. Retico, R. Bellotti, P. Bosco, P. Cerello, G. Gargano, E. Lopez Torres, R. Megna, M. Peccarisi, and M. E. Fantacci, "Combination of computer-aided detection algorithms for automatic lung nodule identification", International Journal of Computer Assisted Radiology and Surgery **7**, 455 (2012).

[10] J. J. Suárez-Cuenca, W. Guo, and Q. Li, "Automated detection of pulmonary nodules in CT: False positive reduction by combining multiple classifiers", in *Proc. SPIE*, Vol. 7963 (2011).

[11] K. Murphy, B. van Ginneken, A. M.R. Schilham, B. J. de Hoop, H. A. Gietema, and M. Prokop, "A large-scale evaluation of automatic pulmonary nodule detection in chest CT using local image features and k-nearest-neighbour classification", Medical Image Analysis **13**, 757 (2009), includes Special Section on the 12th International Conference on Medical Imaging and Computer Assisted Intervention.

[12] Challenge, "LUng Nodule Analysis 2016," `https://luna16.grand-challenge.org/` (2018).

[13] Q. Dou, H. Chen, L. Yu, J. Qin, and P. A. Heng, "Multilevel Contextual 3-D CNNs for False Positive Reduction in Pulmonary Nodule Detection", IEEE Transactions on Biomedical Engineering **64**, (2017).

[14] W. Shen, M. Zhou, F. Yang, D. Dong, C. Yang, Y. Zang, and J. Tian "Learning from Experts: Developing Transferable Deep Features for Patient-Level Lung Cancer Prediction", Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016, 2016 .

[15] A. A. A. Setio, F. Ciompi, G. Litjens, P. Gerke, C. Jacobs, S. J. van Riel, M. M. W. Wille, M. Naqibullah, C. I. Snchez, and B. van Ginneken, "Pulmonary Nodule Detection in CT Images: False Positive Reduction Using Multi-View Convolutional Networks", IEEE Transactions on Medical Imaging **35**, (2016).

[16] W. Li, P. Cao, D. Zhao, and J. Wang, "Pulmonary Nodule Classification with Deep Convolutional Neural Networks on Computed Tomography Images", Computational and Mathematical Methods in Medicine **6215085** (2016).

[17] W. Sun, B. Zheng, and W. Qian "Computer aided lung cancer diagnosis with deep learning algorithms", in *Proc. SPIE*, (2016) .

[18] A. Teramoto, H. Fujita, O. Yamamuro, and T. Tamaki, "Automated detection of pulmonary nodules in PET/CT images: Ensemble false-positive reduction using a convolutional neural network technique", Medical Physics **43**, 2821 (2016).

[19] H. C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep Convolutional Neural Networks for

Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning", IEEE Transactions on Medical Imaging **35**, (2016).

[20] R. Anirudh, J.-J. Thiagarajan, T. Bremer, and H. Kim, "Lung nodule detection using 3d convolutional neural networks trained on weakly labeled data", in *Proc. SPIE*, (2016).

[21] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. van der Laak, B. van Ginneken, and C. I. Snchez, "A survey on deep learning in medical image analysis", Medical Image Analysis **42**, (2017).

[22] F. Ciompi, B. de Hoop, S. J. van Riel, K. Chung, E. T. Scholten, M. Oudkerk, P. A. de Jong, M. Prokop, and B. van Ginneken, "Automatic classification of pulmonary perifissural nodules in computed tomography using an ensemble of 2D views and a convolutional neural network out-of-the-box", Medical Image Analysis **26**, (2015).

[23] B. van Ginneken, A. A. A. Setio, C. Jacobs, and F. Ciompi, "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)* (2015) pp. 286–289.

[24] 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 and Therapy **8**, (2015).

[25] J.-Z. Cheng, D. Ni, Y.-H. Chou, J. Qin, C.-M. Tiu, Y.-C. Chang, C.-S. Huang, D. Shen, and C.-M. Chen, "Computer-Aided Diagnosis with Deep Learning Architecture: Applications to Breast Lesions in US Images and Pulmonary Nodules in CT Scans", Scientific Reports **6** (2016).

[26] W. Shen, M. Zhou, F. Yang, C. Yang, and J. Tian, "Multi-scale Convolutional Neural Networks for Lung Nodule Classification", in *Information Processing in Medical Imaging: 24th International Conference, IPMI 2015, Sabhal Mor Ostaig, Isle of Skye, UK, June 28 - July 3, 2015, Proceedings*, (2015) .

[27] S. Chen, J. Qin, X. Ji, B. Lei, T. Wang, D. Ni, and J. Z. Cheng, "Automatic Scoring of Multiple Semantic Attributes With Multi-Task Feature Leverage: A Study on Pulmonary Nodules in CT Images", IEEE Transactions on Medical Imaging **36**, (2017).

[28] S. J. van Riel, F. Ciompi, M. M. Winkler Wille, A. Dirksen, S. Lam, E. T. Scholten, S. E. Rossi, N. Sverzellati, M. Naqibullah, R. Wittenberg, M. C. Hovinga-de Boer, M. Snoeren, L. Peters-Bax, O. Mets, M. Brink, M. Prokop, C. Schaefer-Prokop, and B. van Ginneken, "Malignancy risk estimation of pulmonary nodules in screening CTs: Comparison between a computer model and human observers", PLOS ONE **12**, 1 (2017).

[29] F. Ciompi, K. Chung, S. J. van Riel, A. A. A. Setio, P. K. Gerke, C. Jacobs, E. T. Scholten, C. Schaefer-Prokop, M. M. W. Wille, A. Marchiano, U. Pastorino, M. Prokop, and B. van Ginneken, "Towards automatic pulmonary nodule management in lung cancer screening with deep learning", Scientific Reports **7** (2017).

[30] A. McWilliams, M. C. Tammemagi, J. R. Mayo, H. Roberts, G. Liu, K. Soghrati, K. Yasufuku, S. Martel, F. Laberge, M. Gingras, S. Atkar-Khattra, C. D. Berg, K. Evans, R. Finley, J. Yee, J. English, P. Nasute, J. Goffin, S. Puksa, L. Stewart, S. Tsai, M. R. Johnston, D. Manos, G. Nicholas, G. D. Goss, J. M. Seely, K. Amjadi, A. Tremblay, P. Burrowes, P. MacEachern, R. Bhatia, M.-S. Tsao, and S. Lam, "Probability of Cancer in Pulmonary Nodules Detected on First Screening CT", New England Journal of Medicine **369**, (2013) .

[31] S. J. van Riel, F. Ciompi, C. Jacobs, M. M. Winkler Wille, E. T. Scholten, M. Naqibullah, S. Lam, M. Prokop, C. Schaefer-Prokop, and B. van Ginneken, "Malignancy risk estimation of screen-detected nodules at baseline CT: comparison of the PanCan model, Lung-RADS and NCCN guidelines", European Radiology **27**, (2017).

[32] "Lung-RADS Version 1.0 Assessment Categories," `https://www.acr.org/~/media/ACR/Documents/PDF/QualitySafety/Resources/LungRADS_AssessmentCategories.pdf`, accessed: 2017-10-25.

[33] "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.

[34] Kaggle competition, "Data science bowl 2017: Can you improve lung cancer detection?" https://www.kaggle.com/c/data-science-bowl-2017 (2017).

[35] F. Liao, M. Liang, Z. Li, X. Hu, and S. Song, "Evaluate the Malignancy of Pulmonary Nodules Using the 3D Deep Leaky Noisy-or Network", arXiv preprint arXiv:1711.08324 (2017).

[36] M. Bergtholdt, R. Wiemker, and T. Klinder, "Pulmonary nodule detection using a cascaded SVM classifier", in *Proc.SPIE*, Vol. 9785 (2016) .

[37] D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, D. P. Naidich, and S. Shetty, "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography", Nature Medicine **25**, (2019).

[38] C. Jacobs and B. van Ginneken, "Googles lung cancer AI: a promising tool that needs further validation", Nature Reviews Clinical Oncology **-**, 1759 (2019).

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition", in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016) pp. 770–778.

[40] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting", Journal of Machine Learning Research **15**, 1929 (2014).

[41] L. Chihara and T. Hesterberg, *Mathematical Statistics with Resampling and R* (Wiley, 2011).

[42] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization", in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (2015).

[43] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition", CoRR **abs/1409.1556** (2014).

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

[45] Frederick J. Larke, Randell L. Kruger, Christopher H. Cagnon, Michael J. Flynn, Michael M. McNitt-Gray, Xizeng Wu, Phillip F. Judy, and Dianna D. Cody, "Estimated Radiation Dose Associated With Low-Dose Chest CT of Average-Size Participants in the National Lung Screening Trial", in *American Journal of Roentgenology* (2011).

[46] S.P. Morozov, A.V. Vladzimirskiy, V.A. Gombolevskiy, V.G. Klyashtorny, I.A. Fedulova, and L.A. Vlasenkov, "Artificial intelligence in lung cancer screening: assessment of the diagnostic accuracy of the algorithm analyzing low-dose computed tomography", in *Tuberculosis and Lung Diseases* (2020)
.

Author contributions

Mr. Trajanovski and Mr. Mavroeidis had full access to the data and can take responsibility for the integrity of the data and the accuracy of the data analysis. Mr. Trajanovski and Mr. Mavroeidis affirm that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; and that any discrepancies from the study as planned (and, if relevant, registered) have been explained.

Concept and design: Trajanovski, Mavroeidis, Leon Swisher, Gebrekidan Gebre

Acquisition, analysis, or interpretation of data: Trajanovski, Mavroeidis, Leon Swisher, Gebrekidan Gebre, Veeling, Wiemker, Klinder, Tahmasebi, Regis, Wald, McKee, Flacke, MacMahon, Pien.

Statistical analysis: Trajanovski, and Mavroeidis.

Drafting of the manuscript: Trajanovski, Mavroeidis, and Leon Swisher

Critical revision of the manuscript for important intellectual content: Trajanovski, Mavroeidis, Leon Swisher, Gebrekidan Gebre, Veeling,

Wiemker, Klinder, Tahmasebi, Regis, Wald, McKee, Flacke, MacMahon, and Pien Supervision: Mavroeidis, and Pien.

Additional information

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

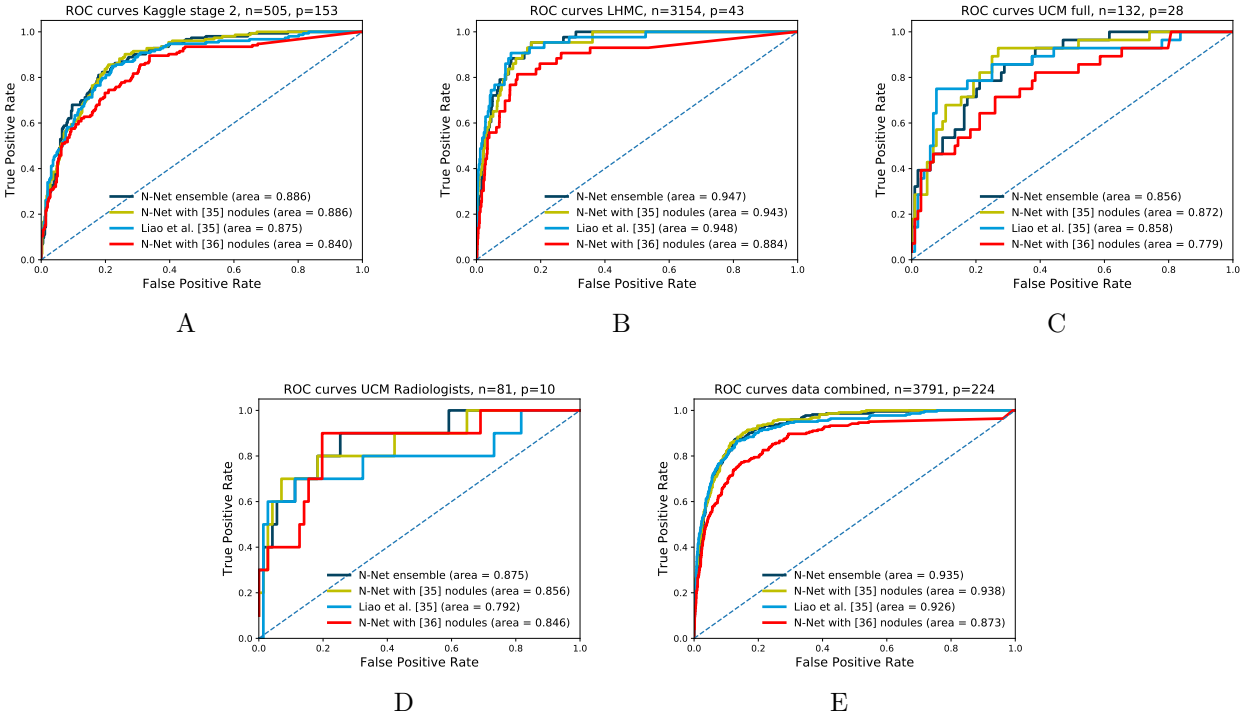


Figure 2: Model Robustness and comparison to state-of-the-art. We use the acronym Nodule-Net (N-Net) to refer to our neural network architecture described in Table 3. Performances are presented for Kaggle stage 2 (505 scans), LHMC (3154 scans), UCM (132 scans), UCM subset with radiologist annotations (81 scans) and an aggregate dataset that combines all the datasets in one graph. The model that employs the *detect_CNN* nodule detector, was trained using 5-fold cross-validation on the NLST data excluding all patients used in the UCM dataset and during inference the average prediction of the 5-folds was used. For the model that employs the *detect_SVM* nodule detector, we have employed again the NLST data for training and we have used three model versions, one trained using 90% of the NLST training data and 0.9 dropout, one using the full NLST data excluding the full UCM patient list (132 patients) and 0.8 dropout and the full NLST data excluding the UCM subset rated by Radiologists (81 patients) and 0.7 dropout. N-net ensemble is simply the average prediction of the N-Net model with *detect_CNN* and the N-Net model with *detect_SVM*. We can make the following observations: (i) The N-net model performs consistently better with the *detect_CNN* nodule detector [35] than with the *detect_SVM* nodule detector [36]; (ii) Model performances are consistent across the different datasets with the performance of the N-Net model with the *detect_CNN* nodule detector ranging between 85.6% and 94.3%; (iii) On the aggregation of all our test sets, we can observe that the AUCs for N-Net and Liao *et al.* are 93.8% and 92.6% respectively. This difference is statistically significant with p-value $P = .0448$. To compute the p-value, we used a two-sided permutation test [41].

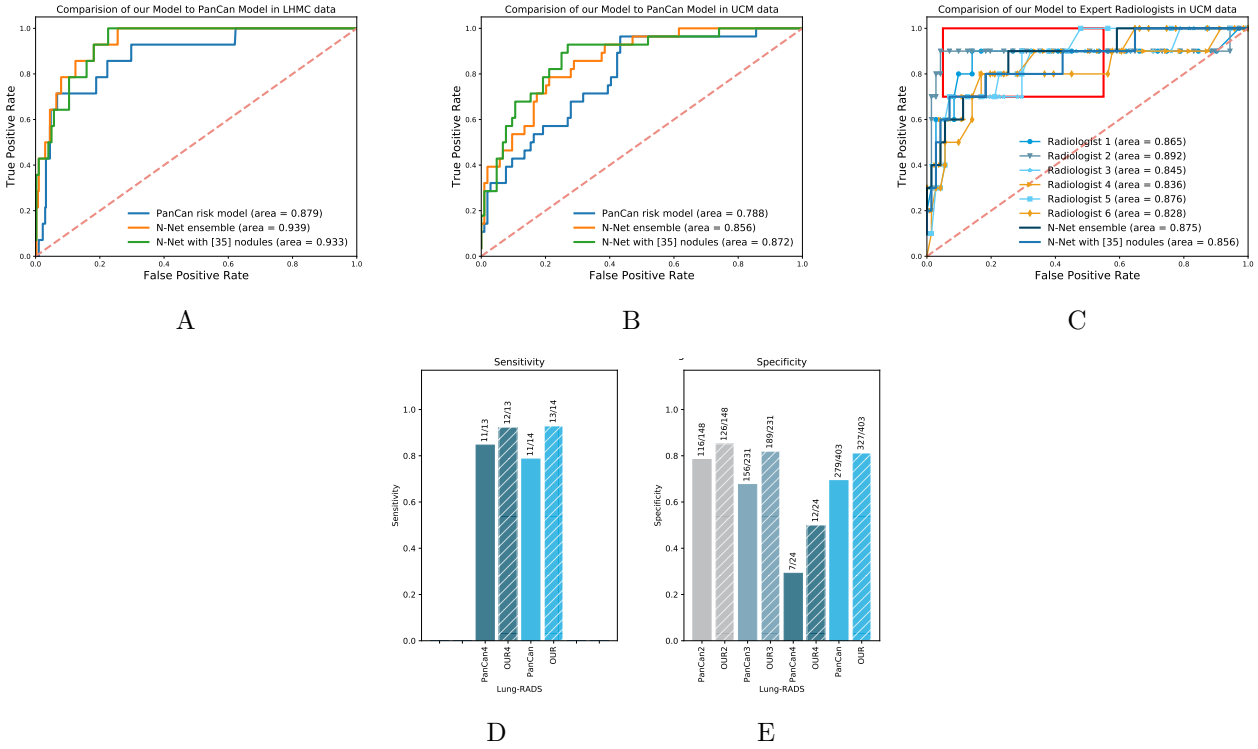


Figure 3: Lung cancer risk assessment performance of our DNN model compared to *PanCan risk model* [30] and radiologists for UCM and LHMC data. A: ROC curve showing the performance of our model and the *PanCan risk model* for LHMC data. For this plot, a subset of the LHMC dataset is used for which we had sufficient information to compute the *PanCan risk model*. This set included 417 scans 14 of which corresponded to cancer cases; B: ROC curve showing the performance of our model and the *PanCan risk model* for all 132 studies in UCM data (with 28 verified cancer cases). C: ROC curve showing the performance of our model compared to radiologists' assessments for 81 studies that have available annotations in UCM data. D: Lung-RADS$^{\text{TM}}$ grouped sensitivity for LHMC data when the specificity is set to 80%. The bars illustrate the number of cancer cases the model is able to identify (true positives), out to the 14 in total cancer cases (13 of which are categorized as Lung-RADS$^{\text{TM}}$ 4); E: Lung-RADS$^{\text{TM}}$ grouped specificity for LHMC data when the sensitivity is set to 93%. "OUR2,3,4" and "PanCan2,3,4" labels refer to the performance achieved for Lung-RADS$^{\text{TM}}$ = 2,3,4 classified cases, while "PanCan" and "OUR" refer to the N-Net model as described in Table 3, trained using the [35] nodules on the NLST dataset excluding the UCM patients. The bars illustrate the number of non-cancer cases the model is able to identify (true negatives), out to the 403 in total non-cancer cases. True negatives are also presented for each Lung-RADS$^{\text{TM}}$ category.

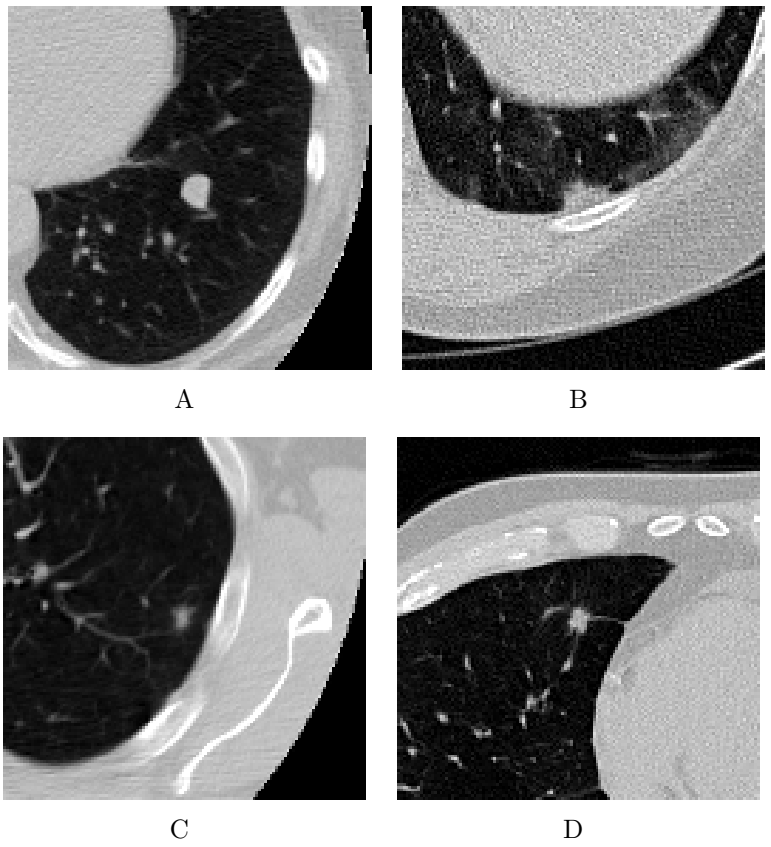
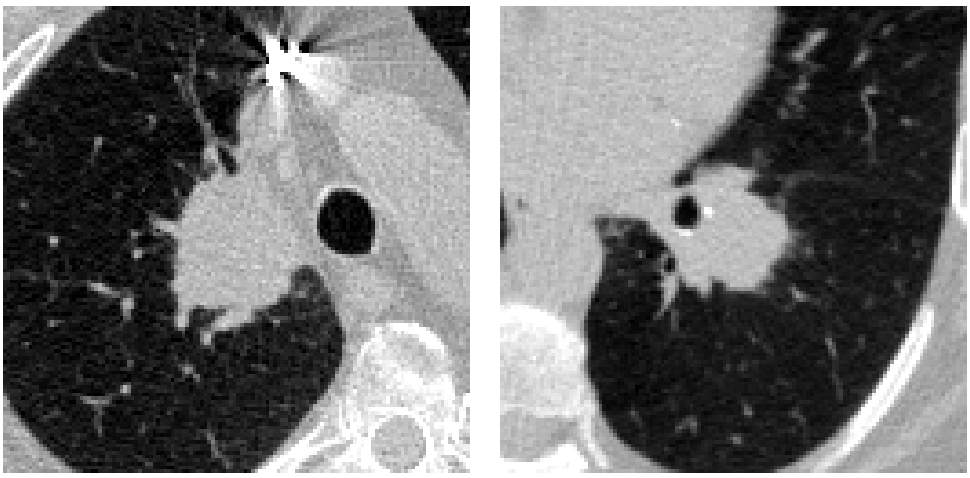


Figure 4: Predictions examples for our DNN model and the *PanCan risk model* [30] in the UCM dataset. A: Example of confirmed malignant nodule: our DNN model assigns a malignancy risk score of 35% while the *PanCan risk model* assigns a malignancy risk score of 5%. Because of the scale difference of the two models, we should note that our DNN model ranks this case at the top 60% high risk cases in the UCM dataset while the *PanCan risk model* ranks this case at the top 38% high risk cases in the UCM dataset. B: Example of benign nodule, our DNN model assigns a malignancy risk score of 10% while the *PanCan risk model* assigns a malignancy risk score of 6.1%. Because of the scale difference of the two models, we should note that our DNN model ranks this case at the top 38% high risk cases in the UCM dataset while the *PanCan risk model* ranks this case at the top 41% high risk cases in the UCM dataset. C: Example of malignant nodule, our DNN model assigns a malignancy risk score of 78% while the *PanCan risk model* assigns a malignancy risk score of 6.4%. Because of the scale difference of the two models, we should note that our DNN model ranks this case at the top 83% high risk cases in the UCM dataset while the *PanCan risk model* ranks this case at the top 42% high risk cases in the UCM dataset. D: Example of malignant nodule, our DNN model assigns a malignancy risk score of 72% while the *PanCan risk model* assigns a malignancy risk score of 4.3%. Because of the scale difference of the two models, we should note that our DNN model ranks this case at the top 80% high risk cases in the UCM dataset while the *PanCan risk model* ranks this case at the top 35% high risk cases in the UCM dataset.



A B

Figure 5: Examples of nodules detected by [35] nodule detector and not by [36]. A: Right upper lobe spiculated mass larger than 4.4 x 3.2 cm. B: Left perihilar mass measuring 3.7 x 2.7 cm.