ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230471

CompLung: Comprehensive Computer-Aided Diagnosis of Lung Cancer

Adam Pardyl[a,b;*], Dawid Rymarczyk[a,b], Joanna Jaworek-Korjakowska[c], Dariusz Kucharski[c], Andrzej Brodzicki[c], Julia Lasek[d], Zofia Schneider[d], Iwona Kucybała[e], Andrzej Urbanik[e], Rafał Obuchowicz[e], Zbisław Tabor[c] and Bartosz Zieliński[a]

[a]Jagiellonian University, Faculty of Mathematics and Computer Science
[b]Jagiellonian University, Doctoral School of Exact and Natural Sciences
[c]AGH University of Science and Technology, Faculty of Electrical Engineering, Automatics, Computer Science, and Biomedical Engineering
[d]AGH University of Science and Technology, Faculty of Geology, Geophysics, and Environmental Protection
[e]Jagiellonian University Medical College, Department of Diagnostic Imaging
ORCiD ID: Adam Pardyl https://orcid.org/0000-0002-3406-6732,
Dawid Rymarczyk https://orcid.org/0000-0002-8543-5200,
Joanna Jaworek-Korjakowska https://orcid.org/0000-0003-0146-8652,
Dariusz Kucharski https://orcid.org/0000-0002-0107-2407,
Andrzej Brodzicki https://orcid.org/0000-0001-7713-526X, Julia Lasek https://orcid.org/0000-0003-2516-1823,
Zofia Schneider https://orcid.org/0000-0003-4987-5361, Iwona Kucybała https://orcid.org/0000-0002-5823-4390,
Andrzej Urbanik https://orcid.org/0000-0001-9541-6727,
Rafał Obuchowicz https://orcid.org/0000-0001-5883-5551, Zbisław Tabor https://orcid.org/0000-0002-9688-9718,
Bartosz Zieliński https://orcid.org/0000-0002-3063-3621

Abstract. Lung cancer is a leading cause of cancer-related deaths, and early diagnosis is crucial for its effective treatment. That is why computer-aided tools have been developed to support particular steps of CT scan analysis, including lung segmentation, suspicious region detection, and patient-level diagnosis. However, none of the previous approaches addressed this process comprehensively. To fill this gap, we introduce CompLung, a comprehensive tool for lung cancer diagnosis that performs all of the above-listed steps in an end-to-end manner. We have trained the CompLung architecture using the publicly available LIDC-IDRI dataset extended with lung segmentation masks obtained from our internal radiologists, which we make publicly available to boost the research on this emerging topic. Finally, we conduct extensive experiments and demonstrate the superior performance and interpretability of CompLung compared to existing methods for lung cancer diagnosis.

1 Introduction

Lung cancer is a leading cause of death in men, and early diagnosis is crucial for its effective treatment [13]. However, analyzing Computed Tomography (CT) scans to detect this disease is a complex and time-consuming process that requires highly trained radiologists. Furthermore, biases and disagreements between doctors usually require consensus, complicating it even more. That is why various machine learning tools were introduced to address those challenges [10, 26, 33, 37].

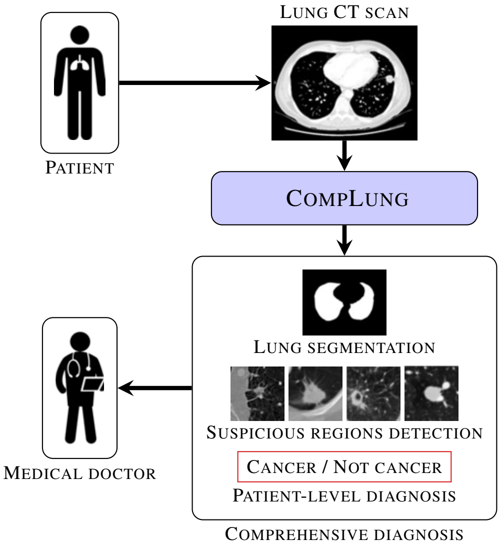


Figure 1: **Overview**: We propose a fully automated pipeline for lung cancer screening. Our CompLung takes as input an unprocessed computer tomography lung scan. The result is an organ segmentation mask, patient-level cancer probability and locations of detected potentially malignant nodules.

The existing approaches are dedicated to automating individual

[*] Corresponding Author. Email: adam.pardyl@doctoral.uj.edu.pl

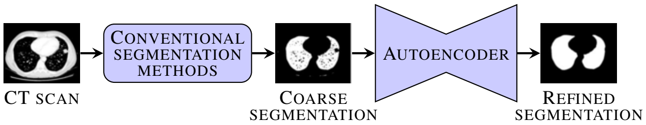


Figure 2: **Lungs segmentation**: In the first stage of CompLung, the lung area is segmented to create a mask for later stages. It is a two-stage process. It uses classical image analysis methods (like thresholding and mathematical morphology) to generate initial segmentation, which is then refined by autoencoder architecture.

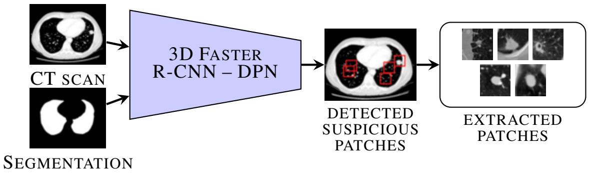


Figure 3: **Suspicious patch identification**: In the second stage, we identify lung regions possibly containing the nodules by adapting 3D Faster R-CNN with Deep 3D Dual Path Net. The further stages consider only those patches.

stages of diagnosis, such as lung segmentation [15, 19] or nodule classification [1, 33]. However, considering those stages individually is unreliable because it does not regard the mistakes made by the previous steps.

Except for the strong research efforts in single steps of lung cancer diagnostics, researchers have introduced a deep learning-based method for the multistep procedure presented in [37]. This approach involves detecting suspicious regions and classifying them to determine if a patient has lung cancer. However, this method has limitations, including the detection of noise outside the lungs as suspicious regions, which can lead to false positives. Additionally, the model's low specificity in detecting suspicious regions may negatively impact patient-level diagnosis as it aims to detect as many cancer regions as possible, which may result in a high number of false positive predictions.

To address these limitations, We introduce CompLung, a comprehensive tool for lung cancer diagnosis that performs all steps of CT scan analysis end-to-end (see Figure 1). It starts with lung segmentation, where the initial mask obtained with classical methods is refined using U-Net [29]. Then, it detects the suspicious regions using a detector based on Faster R-CNN [28]. Those regions are filtered out by region classifier. Finally, the remaining patches are treated as aggregated to obtain patient-level prediction using the Multiple-Instance Learning method [22].

Conducted experiments demonstrate that CompLung overpasses existing approaches on the LIDC-IDRI dataset while being more interpretable. Moreover, it demonstrates an ability to reliably perform multiple diagnosis steps, distinguishing it from other methods. Therefore, it can be considered a reliable diagnostic system for lung cancer diagnosis based on CT scans.

Our contributions can be summarized as follows:

- We propose CompLung, a comprehensive end-to-end diagnostic system for lung cancer diagnosis based on CT scans.
- We extend the LIDC-IDRI dataset with lung segmentations obtained from internal radiologists and make this extension publicly available for other researchers (see Section 4).
- CompLung can be considered a reliable diagnostic system due to its comprehensiveness and high accuracy, achieving over 8% improvement in patient-level classification AUC (area under the receiver operating characteristic curve) score compared to state-of-the-art methods.

2 Related Works

Due to their medical importance, computer-assisted diagnosis methods for lung cancer have been developed for years with various levels of automation [13]. In this section, we first describe existing algorithms for lung organ segmentation. Later we focus on related nodule detection and patient classification methods.

Lung segmentation. The problem of automatic lung segmentation has previously been addressed using conventional methods and deep learning models. [5] uses a region growing supplemented by a reconstruction procedure based on a rolling ball filter for smoothing the segmentation borders. [18] combines thresholding with three-dimensional single-connected components labeling. [16] applies mathematical morphology operations to refine the thresholding results to produce final lung segmentation masks, while [35] uses the Otsu segmentation algorithm for this purpose. [31] adapt optimal thresholding followed by region labeling was also used for lung segmentation. [8] apply minimal graph cutting with Gaussian mixture models (GMMs), while [24] use fuzzy methods.

More recent methods use deep learning to address this problem. [34] applies a 2D convolutional network to produce the segmentation of a 2D scan. Other works use larger models such as Residual U-Nets [9] and Deeplab v3+ [6] for automated lung segmentation [19, 15].

Our CompLung is a hybrid approach, leveraging the low data requirements of conventional segmentation methods with the robustness and flexibility of deep neural networks.

Nodule detection and patient classification. Standard approaches to lung cancer diagnosis do not differentiate between nodule-level and patient-level stages. The patient is considered cancer positive if at least one malignant nodule is detected [13]. Conventional methods generally combine several images transforms to perform nodule detection and manual feature engineering with a conventional classifier for nodule classification. [25] uses a growing neural gas algorithm to segment suspicious areas, followed by a set of transforms to remove blood vessels and bronchi. The detected nodules are then classified by an SVM using shape and texture features. Similar approaches detect nodules using mass-spring models [5] or multistep thresholding [16]. [11] propose an alternative solution, which learns the parameters of conventional image transformations through gradient descent. [1] propose advanced feature engineering, while [23] adapts the channeler ant model.

Recent works focus on the application of deep neural networks to this problem. [10] and [37] use Faster R-CNN [28] to detect nodules and eliminate the need for complex and manually tuned conventional pipelines. At the same time, module classification has been addressed with multiscale convolutional networks [33], Dual Path Networks [37] or 3D ResNet networks [26]. Moreover, [3] proposes a hybrid approach combining conventional image transformation and feature extraction methods with deep neural networks to address the limited availability of training data. Alternatively, weakly supervised multiple instance learning-based methods have been proposed to solve this problem [32, 26]. Most recent models employ self-supervision for representation learning to reduce the manually annotated data needed for training, as the manual labeling process requires trained radiologists and is expensive [12, 26]. Finally, the LUNA16 challenge [30] created a well-defined testing benchmark for algorithms for nodule detection and classification using the publicly available LIDC-IDRI dataset.

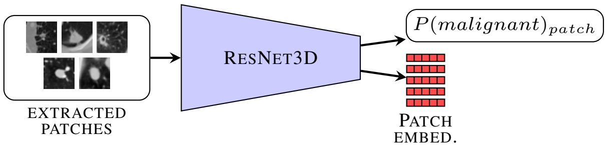


Figure 4: **Patch representation learning**: In the third stage, suspicious patches are used to train a ResNet-18 model that distinguishes between malignant or benign/healthy. Values from its penultimate layer are considered as patch representation.

Despite multiple nodule detection and classification approaches, end-to-end patient-level assessment methods still need to be included. We identify only two modern, deep learning-based works that address the problem of patient-level classification on the LIDC-IDRI dataset. The first of them introduces DeepLung [37], a two-step deep learning algorithm with FASTER R-CNN and 3D-CNN for detection and classification steps, respectively. The second compares weakly-supervised methods for automated lung cancer classification [26].

CompLung extends the deep learning-based nodule detection and classification by adding a patient-level weakly-supervised stage, which improves overall accuracy.

3 CompLung

In this section, we describe CompLung, a comprehensive tool for lung cancer diagnosis working end-to-end. In successive steps described below, it performs lung segmentation, suspicious region detection, patch representation learning, and patient-level diagnosis steps end-to-end.

3.1 Lung segmentation

Our lung segmentation is a two-step method (see Figure 2). The first step uses classical image analysis methods, such as thresholding and mathematical morphology, to obtain an initial segmentation. Because such segmentation may not reflect the anatomical shape of the organ, the second step corrects it using deep learning methods based on autoencoder architecture. We opted for such a hybrid segmentation approach due to a relatively small size of the training dataset [2].

In the first step, we segment the body, and within the body, we segment the lung area, bronchial tree and trachea. The initial lung segmentation is finally obtained by subtracting the mask of the bronchial tree and trachea from the lung area.

The body mask is obtained by applying threshold -191 to each slice (2D view) of the 3D CT image (HU values for soft tissues are usually above this threshold [38]). Then, morphological operations (hole filling and connected component labeling) are applied to find a single connected component related to the body region. Given the segmentation of the body region, we take the region with HU values below -320 as the initial lung area mask. This initial mask has three defects. First, for thick slices, lungs can be merged with intestinal loops. Second, it contains the bronchial tree and the trachea that should be segmented and extracted. Third, the left and right lungs can be merged. The first defect is eliminated by appropriate application of mathematical morphology (erosion followed by filtering the small components and reconstruction of at most two largest components). Eliminating the second defect requires the segmentation of the bronchial tree and the trachea, which starts from finding a slice in which a characteristic pattern of three clusters is visible (two relatively big clusters corresponding to the cross-sections of the two lungs and a relatively small cluster in the middle corresponding to the cross-section of the trachea). Then, a breadth-first search is used starting from the characteristic slice. The third defect is neutralized with the watersheds algorithm that starts from markers obtained after running erosion until two clusters corresponding to the two lungs are obtained. Lung area mask without those defects is returned as the initial lung segmentation.

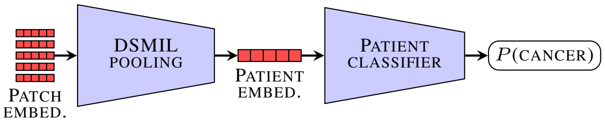


Figure 5: **Patient-level classification**: In the fourth (final) stage, CompLung aggregates representations of the suspicious patches into a single embedding per patient and passes it to the final fully connected layer to obtain a patient-level classification.

In the second step, the initial lung segmentation is corrected using convolutional autoencoder architecture trained with special data augmentations that remove lung parts from the reference segmentation. This way, the model returns more accurate lung segments, even if the classical methods omit some parts.

3.2 Suspicious region identification

In this stage, we identify lung regions containing suspicious changes to limit the operating range of the subsequent steps (see Figure 3). For this purpose, we adopt a nodule candidate detector from [37] based on 3D Faster R-CNN [28] and Deep 3D DPN network [7] trained to detect nodules. The network consists of a U-Net-like encoder-decoder structure and 3D dual path blocks. Segmentation mask created in the previous step is used to remove background. Because a full 3D CT image is too large to fit in GPU memory, we apply a 96^3 sliding window. After detection, we crop 32^3 patches centered on detected nodules for the next stage.

3.3 Patch representation learning

For each patch extracted this way, we want to obtain its representation for the successive aggregation step (see Figure 4). For this purpose, we train the classifier and then use values from its penultimate layer (before the last fully-connected layer) as patch representation (see Figure 4). This classifier is based on ResNet3D-18, a modification of the standard 2D ResNet-18 [14] adapted for 3-dimensional images. It is trained on a regression task, predicting the maximum malignancy score of any nodule occurring in the processed 32^3 patch.

3.4 Patient-level classification

In this stage, we aim to provide a patient-level diagnosis (healthy or requiring further diagnosis) based on the representations of the suspicious patches obtained in the previous step. For this purpose, we use Dual-Stream Multiple Instance Learning (DSMIL) [22], as it demonstrates superior performance compared to other MIL methods. DSMIL assigns weights to all the representations (assigning higher weight for more crucial patches) and then use them to calculate the weighted average, forming a single embedding per patient. This embedding is classified by a fully connected layer, outputting the final patient-level prediction (see Figure 5). By using embedding aggregation instead of aggregating predictions of patches, we achieve higher performance as the final classifier operates on richer representations. Moreover, thanks to limiting the number of patches only to those suspicious, we are able to train this model with less GPU memory.

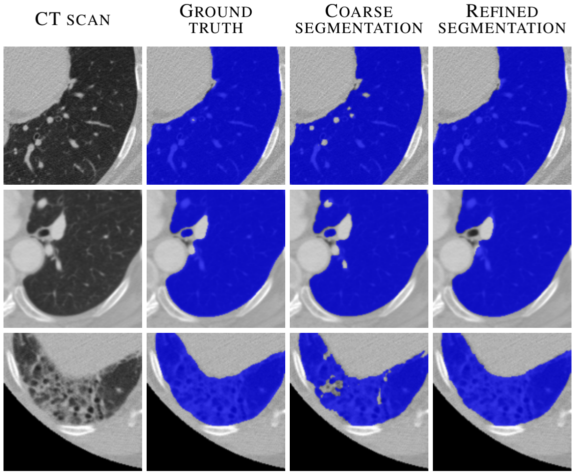


Figure 6: **Lung segmentation**: The first column on the left shows 2D patches sampled from 3D CT scans. The second column presents ground truth masks of lung segmentation prepared by our radiologists. Next, the results of automated lung segmentation created with conventional methods are presented. Finally, the results of autoencoder refinement are shown. One can observe that the autoencoder removes artifacts left by conventional methods.

4 Dataset

In our experiments we use a publicly available lung CT scan dataset described in this section. Additionally, we created a set of segmentation masks for evaluation, which we publish together with this paper.

LIDC-IDRI. We evaluate the proposed approach on the publicly available LIDC-IDRI dataset [2], the largest fully annotated lung CT dataset comprising 1018 scans. We compare our results with the DeepLung [37], MilLung, and AutoLung [26]. Therefore, we use a similar experimental setup, 10-fold cross-validation with the random split. Each experiment divides the dataset into ten equal subsets, where eight subsets are used as the training set, and the remaining two subsets are used for validation and testing, respectively. The experiments are conducted using the first five folds, and the average performance is reported.

Lung segmentation masks. Since the LIDC-IDRI dataset lacks lung segmentation annotations required to evaluate the automated segmentation module, they were created in this project. A team of three radiologists (with 40, 15, and 6 years of experience) individually prepared segmentations using the Slicer 3D [20]. The final segmentation masks correspond to their consensus and are available at https://github.com/gmum/lidc-idri-segment. Some are presented as the ground truth in Figure 6.

5 Experimental setup

In this section, we present evaluation regimes and implementation details used in our experiments.

Nodule malignancy regimes. We use nodule annotations from LIDC-IDRI to train suspicious region identifiers, patch representation encoders, and patient-level classifiers. Moreover, we follow the likelihood of malignancy scale of the dataset, which goes from 0 (highly unlikely) to 5 (highly suspicious) with 3 meaning indeterminate. To account for variations in malignancy scores assigned by the four doctors who annotated the dataset, we take the average of their scores for each nodule. Moreover, nodules with average malignancy scores of 3 are considered neutral and ignored for classification.

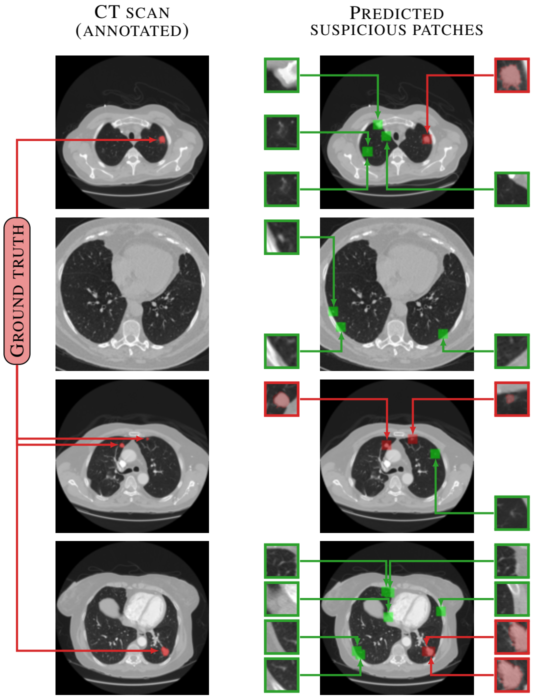


Figure 7: **Suspicious patch identification**: The first column on the left shows 2D slices of 3D CT scans. Nodules annotated by radiologists are marked in red. Next, CT scans with suspicious locations detected by Faster RCNN are presented, true positives (cancerous nodules) in red, and false positives in green. On their sides, magnified slices of extracted 3D patches are shown, with radiologist annotations overlaid in red.

In the experiment named “malignancy > 3”, nodules with scores greater than 3 are classified as positive, and those with scores less than 3 as negative. In the experiment named “malignancy > 0”, all labeled nodules are classified as positive. At the patient level, a scan is labeled as positive if it contains at least one nodule classified as positive.

Implementation. The conventional lung segmentation module was written using OpenCV [4] and scikit-image [36], and we utilized the nnU-Net [17] framework for mask refinement. We changed the nnU-Net base model by eliminating skip connections, resulting in a traditional convolutional autoencoder to minimize the occurrence of high-frequency artifacts. We trained this autoencoder to reconstruct reference segmentations. During the experimentation phase, we used default hyperparameters of the nnU-Net framework. However, to address the common issue of missing cancerous nodules near the segmentation border in conventional segmentation, we incorporated an additional augmentation technique that randomly removes 2-10 elliptical shapes from the images of diameters varying between 2 and 25 voxels.

Patch detection, patch classification, and patient classification were implemented using PyTorch [27]. The Faster-RCNN model was trained using the same hyperparameters and augmentations as specified in [37]. Patch and patient level classifiers are both trained using Adam [21] optimization algorithm. The patch classifier uses mean-squared-error loss to predict the maximum malignancy score of any nodule in the 32^3 voxel patch and a learning rate of 0.0003. Ran-

Table 1: **Patient-level classification:** Results of patient-level classification compared to baseline methods (baseline results taken from [26]). Our CompLung outperforms all baseline methods in both malignancy > 3 and malignancy > 1 regimes.

Regime	Method	Step 1	Step 2	Step 3	AUC
Malignancy > 3	DeepLung	Faster R-CNN	3D DPN		0.83 ± 0.04
	MilLung classification	3D ResNet18	DSMIL		0.77 ± 0.04
		2D ResNet18	DSMIL		0.82 ± 0.02
	MilLung regression	3D ResNet18	DSMIL		0.83 ± 0.03
	AutoLung	Autoencoder	Random Forest		0.80 ± 0.01
	CompLung	Faster R-CNN	3D ResNet18	DSMIL	$\mathbf{0.90 \pm 0.04}$
Malignancy > 0	DeepLung	Faster R-CNN	3D DPN		0.86 ± 0.04
	MilLung classification	3D ResNet18	DSMIL		0.77 ± 0.09
		2D ResNet18	DSMIL		0.84 ± 0.02
	MilLung regression	3D ResNet18	DSMIL		0.82 ± 0.06
	AutoLung	Autoencoder	Random Forest		0.71 ± 0.05
	CompLung	Faster R-CNN	3D ResNet18	DSMIL	$\mathbf{0.93 \pm 0.012}$

Table 2: **Patient-level diagnosis compared to ground-truth obtained from radiologists:** Each row corresponds to predictions obtained based on radiologists' consensus or particular method. Each column corresponds to a ground-true diagnosis considered either for a single doctor or for all of them. Each cell corresponds to the accuracy (%) of a method (row) against the ground truth (column). CompLung achieves the highest accuracy against the radiological consensus over all considered approaches.

	Doctor 1	Doctor 2	Doctor 3	Doctor 4	Average	Consensus
Consensus	81.88%	79.95%	77.78%	61.32%	77.23%	100.00%
DeepLung	72.73%	72.36%	72.10%	65.12%	70.57%	80.65%
MilLung	74.13%	74.52%	70.92%	62.02%	70.45%	78.32%
AutoLung	54.79%	54.79%	55.48%	44.52%	52.40%	73.20%
CompLung	76.43%	70.09%	69.67%	63.00%	72.28%	82.14%

dom flips in all 3D and random affine transform with scale $0.9 - 1.1$ and up to 10 degrees of rotation are used as augmentation. Training examples are randomly sampled with class balancing.

For the patient-level DSMIL classifier, we use standard cross-entropy loss. Grid search is applied to find the best hyperparameters, picking the learning rate from $[1 \cdot 10^{-3}, 2 \cdot 10^{-4}, 1 \cdot 10^{-4}, 5 \cdot 10^{-5}]$, weight decay from $[5 \cdot 10^{-3}, 1 \cdot 10^{-4}, 1 \cdot 10^{-5}]$, a dropout from $[0, 0.1, 0.2]$, and the number of add-on nonlinear layers for DSMIL between $[0, 1]$.

All models used in this project can be trained and evaluated using a single NVIDIA V100 GPU. An end-to-end inference pass takes around 20 minutes per batch using our experimental implementation.

Baselines. We compare our CompLung to other state-of-the-art lung cancer screening methods. The first baseline method is DeepLung [37], a dual-stage detector-classifier method trained in a fully-supervised regime. Next, we compare to MilLung [26], a data-efficient multiple instance learning-based patient-level classification approach. Finally, we consider AutoLung [26], a weakly-supervised method built upon anomaly detection using an autoencoder trained to reconstruct healthy lung scans and a random forest classifier for detecting anomalies in the autoencoder generated representations.

6 Results

In the experiments, we first compare our CompLung to comparable methods in patient-level diagnosis tasks. Then, we analyze the performance of each step of CompLung pipeline and provide qualitative examples of their work.

6.1 Patient screening

The main goal of CompLung is patient-level diagnosis. We evaluate its performance in this task and compare it against state-of-the-art baseline methods. In Table 1, we present AUC score evaluation, where baseline scores are taken from [26]. CompLung achieves superior performance compared to all three state-of-the-art baseline methods for both malignancy > 3 and malignancy > 0 regimes, scoring over 8% better AUC score than the second-best DeepLung (0.90 to 0.83 for malignancy > 3 and 0.93 to 0.86 for malignancy > 0). This improvement indicates the diagnostic effectiveness of our method.

Following [37] and [26] we present a comparison between algorithms and experienced doctors[1]. In Table 2 accuracy results for malignancy > 3 regimes are shown. The baseline scores are taken from [26]. The accuracy is calculated only on patients from the test set with nodules annotated by four radiologists. The average agreement between particular doctors and a consensus (average over scores of four radiologists) is provided. One can observe that our CompLung outperforms, on average, all baseline methods and one of the doctors.

[1] Radiologists annotations are taken from https://github.com/wentaozhu/DeepLung/blob/master/nodcls/annotationdetclssgm_doctor.csv

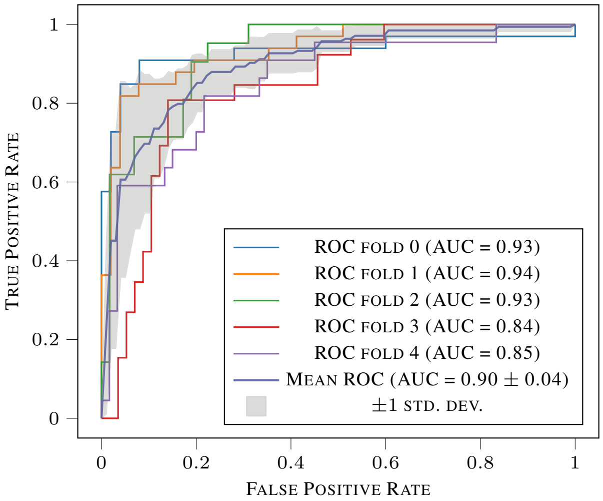


Figure 8: **Patient-level classification**: ROC curves for patient-level classification in the malignancy > 3 scenario. Area under curve (AUC) scores for each fold are presented in the plot legend. The mean ROC curve and ±1 standard deviation margin are marked on the plot with blue and gray respectively.

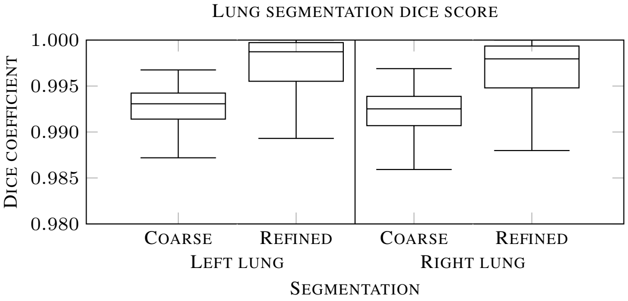


Figure 9: **Lung segmentation**: Box plots of the dice coefficient for left and right lung segmentations before and after refinement. The use of the autoencoder refinement stage significantly improves the dice score. The resultant refined segmentation map is almost identical to the ground truth annotations.

Additionally, we provide Receiver Operating Characteristic (ROC) curves for each cross-validation test fold in malignancy > 3 regimes, shown in Figure 8. Considering the small size of each test set (around 90 items), we conclude that our method performs well for each test fold.

6.2 Pipeline analysis

In this section, we show the evaluation results of each step of CompLung pipeline.

Lung segmentation results. In this experiment, we demonstrate that our automated lung segmentation process can separate lung regions precisely enough for further processing. We compare the coarse segmentation maps computed only using conventional methods and refined maps processed with an autoencoder to reference segmentations created by radiologists. In Figure 9, we present box plots of dice coefficient scores for left and right lung segmentation before and after refinement. The average dice score for both lungs after refinement is 0.99. Therefore, we conclude that automated lung segmentations are almost identical to the reference annotations created by radiologists. In comparison, a pure U-Net segmentation model we trained achieved a DICE score of only 0.95. We hypothesises this was caused by a relatively small size of the training dataset. Figure 6 shows the example masks generated in this step.

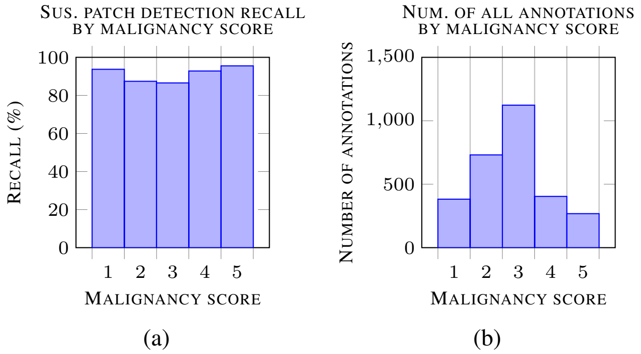


Figure 10: **Sensitivity of detecting suspicious patches**: Plot a) presents the recall of suspicious patch detection aggregated by the maximum malignancy of nodules provided by radiologists. Plot b) shows the number of all annotations matching the given malignancy score in the test set. Note that we do not average the malignancy score across doctors in this study but rather take the maximum value. We observe that nodules annotated with high malignancy scores are detected with high recall even though they are underrepresented in the training set.

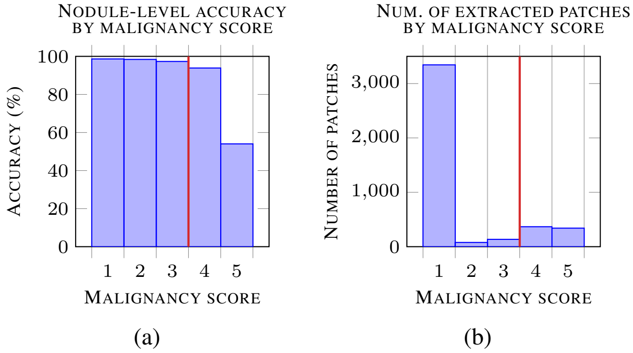


Figure 11: **Patch classification accuracy**: Plot a) presents the accuracy of nodule level classification for malignancy > 3 regime, aggregated by the average malignancy score of the given nodule. Plot b) shows the number of all test patches matching the given malignancy score. The red line denotes the classification threshold. One can observe that the patch classifier generates more false negatives than false positives due to class imbalance.

Suspicious patches identification results. Moreover, we analyze the performance of the suspicious patch identification stage in detecting nodules annotated by radiologists. We calculate the percentage of nodules (regardless of annotated malignancy score) detected as suspicious areas, i.e. the recall rate. The results aggregated by annotation malignancy score are presented in Figure 10a. As a reference, Figure 10b shows the total number of annotations by malignancy score. Overall, almost all nodules are detected, and nodules with a high malignancy score (over 3) are picked the most often. The examples of suspicious locations and patches extracted from them are shown in Figure 7.

Patch classification results. Next, we evaluate patch classification accuracy based on the extracted suspicious patches. The accuracy of patch-level classification in the malignancy > 3 regime aggregated by the patch malignancy is shown in Figure 11a. As a reference, Fig-

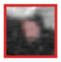
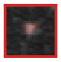
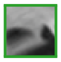
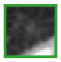
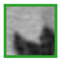
	Sus. patch	Patch label	Patch pred.	DSMIL att. ↓	Patient pred.	Patient label.
Patient A		3.5	3.873	0.523	0.606 (cancer)	cancer
		3.25	3.608	0.402		
		0	0.431	0.014		
		0	0.340	0.012		
		0	0.105	0.011		
	⋮	⋮	⋮	⋮		
Patient B		0	3.267	0.238	0.254 (healthy)	healthy
		0	2.949	0.203		
		0	1.537	0.042		
		0	0.729	0.024		
		0	0.644	0.022		
	⋮	⋮	⋮	⋮		

Figure 12: **Patch classification vs. patient classification**: Visualization of patch classification and patient classification for two sampled CT scans. The first column for each patient presents five extracted suspicious patches ordered by the DSMIL attention scores. Radiologist's annotations are overlaid in red. The second column shows the average scores assigned by four radiologists to each patch. The third column shows the result of the patch-level classifier, and the fourth one corresponds to the attention score of the DSMIL aggregation. The two final columns present the patient-level prediction and the ground truth label. Observe that using DSMIL improves accuracy compared to using the maximum patch scores as the patient score because, in the case of Patient B, the max patch score incorrectly indicates a positive diagnosis. In contrast, the final patient score obtained by DSMIL is correctly negative.

ure 11b shows the total number of extracted patches by malignancy score. We observe that patch-level classification correctly removes most false positive patches extracted by the previous stage but fails to recognize some true cancerous nodules. Note that the final patient-level classifier does not use this score but a latent representation of each patch. Example patch level evaluation is presented in the left part of Figure 12.

Patient classification results. Finally, we analyze the patient-level classification accuracy based on latent representations acquired in the previous step. The accuracy at the patient-level in the malignancy > 3 regime aggregated by the maximum malignancy of any nodule on the scan is shown in Figure 13a. Figure 13b shows the number of patients by malignancy score as a reference. One can observe that the accuracy is high for all malignancy scores above 3 and those below 1, while it is lower for scores just below 3. However, it is expected because CompLung is designed for patient screening. Therefore, false positive results are preferred to false negative ones. Sample patient-level diagnosis is presented in the right part of Figure 12.

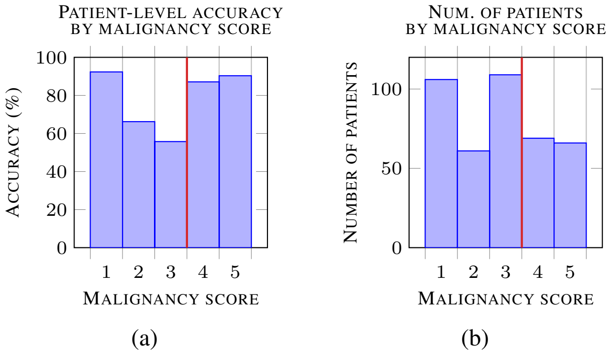


Figure 13: **Patient classification accuracy**: Plot a) presents the accuracy of patient-level classification for malignancy > 3 regime, aggregated by the maximum malignancy score of any nodule in the patient CT scan as calculated based on the consensus of all radiologists. Plot b) shows the number of all scans matching the given malignancy score in the test set. The red line denotes the malignancy threshold. We observe that scans with nodules just under the threshold are more prone to misclassification, while malignant cases are classified well.

7 Conclusions

In this paper, we proposed CompLung, a comprehensive end-to-end tool for lung cancer diagnosis that performs all steps of CT scan analysis. The system integrates state-of-the-art techniques, including U-Net and Faster R-CNN, to refine lung segmentation and detect suspicious regions. Our contributions also include the extension of the LIDC-IDRI dataset with lung segmentations obtained from our radiologists, which we make publicly available for other researchers.

The results of the experiments unequivocally demonstrate that CompLung surpasses existing approaches on the LIDC-IDRI dataset while offering superior interpretability. Consequently, by comprehensively and precisely addressing the challenges of lung cancer diagnosis, CompLung can positively impact patient outcomes and relieve healthcare professionals of some of their burdens. A potential avenue in improving the tool would be training a foundation model for generic CT scan understanding and fine-tuning on LIDC-IDRI, which would help to offset the relatively small size of the dataset.

Acknowledgements

Acknowledge support of National Center for Research and Development (NCBR, Poland) under grant no. POIR.01.01.01-00-1666/20. We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Centers: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2021/014764 and PLG/2023/016441.

References

[1] Hugo JWL Aerts, Emmanuel Rios Velazquez, Ralph TH Leijenaar, et al., 'Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach', *Nature communications*, **5**(1), 1–9, (2014).

[2] Samuel G Armato III, Geoffrey McLennan, Luc Bidaut, et al., 'The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans', *Medical physics*, **38**(2), 915–931, (2011).

[3] A Asuntha and Andy Srinivasan, 'Deep learning for lung cancer detection and classification', *Multimedia Tools and Applications*, **79**(11), 7731–7762, (2020).

[4] G. Bradski, 'The OpenCV Library', *Dr. Dobb's Journal of Software Tools*, (2000).

[5] Donato Cascio, Rosario Magro, Francesco Fauci, Marius Iacomi, and Giuseppe Raso, 'Automatic detection of lung nodules in ct datasets based on stable 3d mass–spring models', *Computers in biology and medicine*, **42**(11), 1098–1109, (2012).

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, 'Encoder-decoder with atrous separable convolution for semantic image segmentation', in *ECCV*, (2018).

[7] Yunpeng Chen, Jianan Li, Huaxin Xiao, Xiaojie Jin, Shuicheng Yan, and Jiashi Feng, 'Dual path networks', *NeurIPS*, **30**, (2017).

[8] Shuangfeng Dai, Ke Lu, Jiyang Dong, Yifei Zhang, and Yong Chen, 'A novel approach of lung segmentation on chest ct images using graph cuts', *Neurocomputing*, **168**, 799–807, (2015).

[9] Foivos I Diakogiannis, François Waldner, Peter Caccetta, and Chen Wu, 'Resunet-a: A deep learning framework for semantic segmentation of remotely sensed data', *ISPRS Journal of Photogrammetry and Remote Sensing*, **162**, 94–114, (2020).

[10] Jia Ding, Aoxue Li, Zhiqiang Hu, and Liwei Wang, 'Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks', in *MICCAI*, pp. 559–567. Springer, (2017).

[11] Amal A Farag, Hossam E Abd El Munim, James H Graham, and Aly A Farag, 'A novel approach for lung nodules segmentation in chest ct using level sets', *IEEE Transactions on Image Processing*, **22**(12), 5202–5213, (2013).

[12] Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Zongwei Zhou, Michael B Gotway, and Jianming Liang, 'Learning semantics-enriched representation via self-discovery, self-classification, and self-restoration', in *MICCAI*, pp. 137–147. Springer, (2020).

[13] Amitava Halder, Debangshu Dey, and Anup K Sadhu, 'Lung nodule detection from feature engineering to deep learning in thoracic ct images: a comprehensive review', *Journal of digital imaging*, **33**(3), 655–677, (2020).

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, 'Deep residual learning for image recognition', in *CVPR*, pp. 770–778, (2016).

[15] Johannes Hofmanninger, Forian Prayer, Jeanny Pan, Sebastian Röhrich, Helmut Prosch, and Georg Langs, 'Automatic lung segmentation in routine imaging is primarily a data diversity problem, not a methodology problem', *European Radiology Experimental*, **4**(1), 1–13, (2020).

[16] Saleem Iqbal, Khalid Iqbal, Fahim Arif, Arslan Shaukat, Aasia Khanum, et al., 'Potential lung nodules identification for characterization by variable multistep threshold and shape indices from ct images', *Computational and mathematical methods in medicine*, **2014**, (2014).

[17] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein, 'nnu-net: a self-configuring method for deep learning-based biomedical image segmentation', *Nature methods*, **18**(2), 203–211, (2021).

[18] Hoon-seok Jang, Wook-Jin Choi, and Tae-Sun Choi, 'Optimal fuzzy rule based pulmonary nodule detection', *Adv Sci Technol Lett*, **29**, 75–8, (2013).

[19] Anita Khanna, Narendra D Londhe, S Gupta, and Ashish Semwal, 'A deep residual u-net convolutional neural network for automated lung segmentation in computed tomography images', *Biocybernetics and Biomedical Engineering*, **40**(3), 1314–1327, (2020).

[20] Ron Kikinis, Steve D Pieper, and Kirby G Vosburgh, '3d slicer: a platform for subject-specific image analysis, visualization, and clinical support', in *Intraoperative imaging and image-guided therapy*, 277–289, Springer, (2013).

[21] Diederik P Kingma and Jimmy Ba, 'Adam: A method for stochastic optimization', *arXiv preprint arXiv:1412.6980*, (2014).

[22] Bin Li, Yin Li, and Kevin W Eliceiri, 'Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning', in *CVPR*, pp. 14318–14328, (2021).

[23] E Lopez Torres, Elisa Fiorina, Francesco Pennazio, et al., 'Large scale validation of the m5l lung cad on heterogeneous ct datasets', *Medical physics*, **42**(4), 1477–1489, (2015).

[24] T Manikandan and N Bharathi, 'Lung cancer detection using fuzzy auto-seed cluster means morphological segmentation and svm classifier', *Journal of medical systems*, **40**, 1–9, (2016).

[25] Stelmo Magalhães Barros Netto, Aristófanes Corrêa Silva, Rodolfo Acatauassú Nunes, and Marcelo Gattass, 'Automatic segmentation of lung nodules with growing neural gas and support vector machine', *Computers in biology and medicine*, **42**(11), 1110–1121, (2012).

[26] Adam Pardyl, Dawid Rymarczyk, Zbisław Tabor, and Bartosz Zieliński, 'Automating patient-level lung cancer diagnosis in different data regimes', in *ICONIP*. Springer International Publishing, (2022).

[27] Adam Paszke, Sam Gross, Francisco Massa, et al., 'Pytorch: An imperative style, high-performance deep learning library', *NeurIPS*, **32**, (2019).

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, 'Faster r-cnn: Towards real-time object detection with region proposal networks', *NeurIPS*, **28**, (2015).

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, 'U-net: Convolutional networks for biomedical image segmentation', in *MICCAI*, pp. 234–241. Springer, (2015).

[30] Arnaud Arindra Adiyoso Setio, Alberto Traverso, Thomas De Bel, et al., 'Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the luna16 challenge', *Medical image analysis*, **42**, 1–13, (2017).

[31] Furqan Shaukat, Gulistan Raja, Ali Gooya, and Alejandro F Frangi, 'Fully automatic detection of lung nodules in ct images using a hybrid feature set', *Medical physics*, **44**(7), 3615–3629, (2017).

[32] Wei Shen, Mu Zhou, Feng Yang, Di Dong, Caiyun Yang, Yali Zang, and Jie Tian, 'Learning from experts: Developing transferable deep features for patient-level lung cancer prediction', in *MICCAI*, pp. 124–131. Springer, (2016).

[33] Wei Shen, Mu Zhou, Feng Yang, Caiyun Yang, and Jie Tian, 'Multi-scale convolutional neural networks for lung nodule classification', in *IPMI*, pp. 588–599. Springer, (2015).

[34] Johnatan Carvalho Souza, João Otávio Bandeira Diniz, Jonnison Lima Ferreira, et al., 'An automatic method for lung segmentation and reconstruction in chest x-ray using deep neural networks', *Computer methods and programs in biomedicine*, **177**, 285–296, (2019).

[35] Joel Chia Ming Than, Norliza Mohd Noor, Omar Mohd Rijal, Ashari Yunus, and Rosminah Md Kassim, 'Lung segmentation for hrct thorax images using radon transform and accumulating pixel width', in *2014 IEEE Region 10 Symposium*, pp. 157–161. IEEE, (2014).

[36] Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, et al., 'scikit-image: image processing in python', *PeerJ*, **2**, e453, (jun 2014).

[37] Wentao Zhu, Chaochun Liu, Wei Fan, and Xiaohui Xie, 'Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification', in *WACV*, pp. 673–681. IEEE, (2018).

[38] Shabana R Ziyad, Venkatachalam Radha, and Thavavel Vayyapuri, 'Overview of computer aided detection and computer aided diagnosis systems for lung nodule detection in computed tomography', *Current Medical Imaging*, **16**(1), 16–26, (2020).