Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

LNDb Challenge on automatic lung cancer patient management

João Pedrosa[a,*], Guilherma Aresta[a,b], Carlos Ferreira[a,b], Gurraj Atwal[c], Hady Ahmady Phoulady[c], Xiaoyu Chen[d], Rongzhen Chen[d], Jiaoliang Li[d], Liansheng Wang[d], Adrian Galdran[e], Hamid Bouchachia[e], Krishna Chaitanya Kaluva[f], Kiran Vaidhya[f], Abhijith Chunduru[f], Sambit Tarai[f], Sai Prasad Pranav Nadimpalli[f], Suthirth Vaidhya[f], Ildoo Kim[g], Alexandr Rassadin[h], Zhenhuan Tian[i], Zhongwei Sun[j], Yizhuan Jia[j], Xuejun Men[j], Isabel Ramos[a,k], António Cunha[a,l], Aurélio Campilho[a,b]

[a] *Institute for Systems and Computer Engineering, Technology and Science (INESC TEC), Porto, Portugal*
[b] *Faculty of Engineering of the University of Porto (FEUP), Porto, Portugal*
[c] *Department of Computer Science, California State University, Sacramento, USA*
[d] *Department of Computer Science, School of Informatics, Xiamen University, China*
[e] *Department of Computing and Informatics, Bournemouth University, UK*
[f] *Predible Health, Bangalore, India*
[g] *Kakao Brain, Seongnam-si, South Korea*
[h] *xperience.ai, Nizhny Novgorod, Russia*
[i] *Department of Thoracic Surgery, Peking Union Medical College Hospital, Peking Union Medical College, Beijing, China*
[j] *Mediclouds Medical Technology, Beijing, China*
[k] *Department of Radiology, Centro Hospitalar e Universitário de S. João, Porto, Portugal*
[l] *University of Trás-os-Montes e Alto Douro (UTAD), Vila Real, Portugal*

ARTICLE INFO

Article history:
Received September 9, 2020
Received in final form September 9, 2020
Accepted September 9, 2020
Available online September 9, 2020

Communicated by —

Keywords: lung cancer, low dose computed tomography, pulmonary nodules, follow-up recommendation, deep learning, challenge.

ABSTRACT

Lung cancer is the deadliest type of cancer worldwide and late detection is the major factor for the low survival rate of patients. Low dose computed tomography has been suggested as a potential screening tool but manual screening is costly, time-consuming and prone to variability. This has fuelled the development of automatic methods for the detection, segmentation and characterisation of pulmonary nodules. In spite of promising results, the application of automatic methods to clinical routine is not straightforward and only a limited number of studies have addressed the problem in a holistic way, using automatic methods to obtain patient follow-up recommendations from the computed tomography image. With the goal of advancing the state of the art in lung cancer medical image analysis, the Lung Nodule Database (LNDb) Challenge on automatic lung cancer patient management was organized. The LNDb Challenge addressed lung nodule detection, segmentation and characterization as well as prediction of patient follow-up according to the 2017 Fleischner society pulmonary nodule guidelines. 294 CT scans were thus collected retrospectively at the Centro Hospitalar e Universitário de São João in Porto, Portugal and each CT was annotated by at least one radiologist. Annotations comprised nodule centroids, segmentations and subjective characterization. A total of 947 participants registered for the challenge and 11 successful submissions for at least one of the sub-challenges were received. All submitted methods relied, at least partly, in deep learning methodologies. A maximum nodule detection sensitivity below 0.4 (and 0.7) for nodules indentified by at least one (and two) radiologist(s) at 1 false positive per scan was obtained, remaining the most challenging task in lung cancer image analysis. For nodule segmentation, a maximum Jaccard score of 0.567 was obtained, surpassing the interobserver variability. In terms of nodule texture characterization, a maximum Fleiss-Cohen weighted Cohen's kappa of 0.733 was obtained, with part solid nodules being particularly challenging to classify correctly. For patient follow-up prediction, a maximum Fleiss-Cohen weighted Cohen's kappa of 0.580 was obtained. Detailed analysis of the proposed methods and the differences in performance allow to identify the major challenges remaining and future directions. The LNDb Challenge and associated dataset remain publicly available so that future methods can be tested and benchmarked, promoting the development of new algorithms in lung cancer medical image analysis and patient follow-up recommendation.

1. Introduction

Lung cancer is the deadliest type of cancer worldwide for both men and women (Siegel et al., 2019). Though changes in the smoking patterns in the general population have been largely responsible for decreasing trends in incidence and mortality rates in recent decades, lung cancer is still responsible for over double the cancer deaths of colorectal cancer, the second deadliest cancer type, and is projected to remain the deadliest type of cancer in the near future. Progress in increasing lung cancer survival rate has also been notoriously slow in contrast to other cancer types, mainly due to late diagnosis of the disease. Low-dose computed tomography (CT) has long been suggested as a potential early screening tool and a 20% reduction in lung cancer mortality has been demonstrated for lung cancer risk groups (The National Lung Screening Trial Research Team, 2011). Nevertheless, translation of these screening programs to the general population has been challenging due to equipment and personnel costs and the complexity of the task. Namely, lung nodules present a large range of shapes and characteristics and thus the identification and characterization of these abnormalities is not trivial and prone to high interobserver variability. Computer-aided diagnosis (CAD) systems can thus facilitate the adoption and generalization of screening programs by reducing the burden on the clinicians and providing a second-opinion.

Extensive research has been conducted on the development of CAD systems for lung cancer screening, typically divided in three main tasks: pulmonary nodule detection, nodule segmentation and nodule characterization and/or classification.

The task of nodule detection aims to automatically find all nodules present in a CT scan and is likely the most active field of research in lung cancer screening/management. Early studies typically followed a pipeline consisting of preprocessing through lung and airways/vessel segmentation, followed by nodule candidate detection through gray-level thresholding and spatial/geometric features and a final false positive reduction step based on feature extraction from each nodule candidate and the use of fixed rules or supervised learning such as support vector machines (SVM) or neural networks (NN) for the classification of each nodule candidate (Reeves and Kostis, 2000; Han et al., 2014; Messay et al., 2010; Murphy et al., 2009). Recently, deep learning based methods have shown especially promising results. One of the first works showing the advantages of deep learning for nodule detection was by Golan et al. (2016), where the authors proposed a patch-based convolutional neural network (CNN), achieving a sensitivity of 0.712 at 10 false positives (FP) per scan, and outperforming previous methods. The LUNA16 nodule detection challenge has highlighted the advantages of these methodologies, as most of the best performing methods relied on deep learning (Setio et al., 2017). Ding et al. (2017) used a region-based CNN for candidate detection which received as input three adjacent axial slices, followed by a false positive reduction step using a 3D CNN, obtaining a sensitivity above 0.90 at 1 FP per scan, one of the highest scores on LUNA16. In Aresta et al. (2018), the YOLOv2 architecture was used to perform lung nodule detection on CT axial slices achieving a sensitivity of 0.926 at 0.25 FPs per scan for nodules with diameter >4mm.

Nodule segmentation aims to automatically segment the nodule borders so that nodule volume can be obtained, an important factor in patient management and follow-up. As in nodule detection, most recent literature uses deep learning architectures for segmentation, extracting a 2D or 3D region around the nodule centroid and performing segmentation in that region of interest. A modified 3D U-Net for joint segmentation and malignancy prediction was proposed by Wu et al. (2018), obtaining a Dice of 0.7405. Given the difficulty of training 3D CNNs, Wang et al. (2017) proposed a central-focused CNN combining 2D and 2.5D features and obtaining a Dice of 0.8215. The differences in nodule appearance is one of the main challenges for robust nodule segmentation as sub-solid and ground-glass opacities (GGO) present lower constrast in relation to lung parenchyma, which leads to decreased performance (Aresta et al., 2019).

*Corresponding author: Email: joao.m.pedrosa@inesctec.pt

Nodule characterization or classification aims to predict further clinically relevant information regarding the nodule. Most studies focus on the classification of nodules regarding their malignancy. Earlier studies were based on feature extraction followed by classification, such as in Gonçalves et al. (2018), where 297 features (shape-, intensity- and texture-based) are computed and an SVM classifier is used, obtaining an area under the curve (AUC) of 0.962. The use of deep learning architectures has allowed for even higher performance, with reported AUC of 0.993 by Causey et al. (2018) using a 3D CNN to extract features, followed by a random forest classifier. The characterization of the nodule according to features related to malignancy has also received some attention, the most studied of which is texture which is often divided in three classes: solid, sub-solid and GGO. Ferreira et al. (2018) proposed a 2.5D CNN for nodule texture classification obtaining an accuracy of 0.833. The prediction of nodule features such as texture as an intermediate step to increase robustness and interpretability of malignancy classification has recently received attention. In Shen et al. (2019), a 3D CNN with two outputs is proposed: the first corresponds to low-level nodule features (texture, calcification, margin, sphericity and subtlety), whereas the second output corresponds to malignancy and is dependent on the features used for the prediction of the low-level nodule features.

Given the dependence of deep learning methods on large datasets with robust ground truth, the publication of annotated datasets has been a hugely important contribution for the community. The most widely used public database for lung cancer medical image analysis research is the LIDC-IDRI (Armato et al., 2011), which contains 1018 CT scans, each annotated by four radiologists. Annotations comprise nodule segmentation and subjective characterization (McNitt-Gray et al., 2007), making this an extremely useful database for the development of CAD approaches in lung cancer. The NLST database is also widely recognised and contains CTs from 26.722 patients, though nodule segmentation and characterization are not available and nodule position is limited to the slice where a nodule was found (The National Lung Screening Trial Research Team, 2011).

Nevertheless, adoption of CAD systems in clinical practice is not straightforward. In spite of the promising results in literature in the independent tasks of nodule detection, segmentation and classification, the final goal of patient management and follow-up, which is dependent on the above tasks, has received little attention. The most widely used guidelines for patient management in the case of incidental nodule findings are the 2017 Fleischner society pulmonary nodule guidelines (MacMahon et al., 2017), which give a recommendation on patient follow-up depending on the number of nodules found, their sizes and texture and patient risk. Following such recommendations allow radiologists to make knowledge-based decisions regarding nodule findings for better patient management while reducing the number of unnecessary follow-up examinations. As such, the design of an impactful CAD system should take these guidelines into account in order to be fully integrated into clinical practice.

The goal of the Lung Nodule Database (LNDb) Challenge on automatic lung cancer patient management was thus to establish a common online and public database and benchmarking framework for this task. More precisely, the LNDb challenge aimed to evaluate and compare the performance of several approaches in the automatic classification of CT scans according to the 2017 Fleischner society pulmonary nodule guidelines. The performance of different methods for the intermediate tasks required for patient classification according to the Fleischner guidelines - nodule detection, segmentation and texture characterization - was also evaluated.

2. Challenge Description

2.1. Overview

The LNDb challenge was made up of a main challenge and three sub-challenges related to the automatic classification of CT scans according to the 2017 Fleischner society pulmonary nodule guidelines for patient follow-up recommendation:

- Main Challenge - Fleischner Classification;
- Challenge A - Nodule Detection;
- Challenge B - Nodule Segmentation;

- Challenge C - Nodule Texture Characterization.

Participants could choose whether to participante only in the main challenge, in a single or multiple challenges or in all challenges.

The LNDb Challenge was hosted on Grand Challenge[1], a well-known challenge platform, which allows for an easy setup of the platform and online evaluation of results. To further promote the challenge, LNDb was organized in conjunction with the 17th International Conference on Image Analysis and Recognition (ICIAR 2020) and monetary prizes were awarded to the top scorers in each of the challenge tasks upon submission and acceptance to ICIAR 2020.

The challenge was held between the 20th November 2019 and the 20th February 2020 and was composed of two evaluation stages: train/validation and test. On the train/validation, participants were given access to the training data and annotations, allowing for the training and refinement of their algorithms. Participants were also encouraged to submit their results for online evaluation using a cross-validation scheme (cf. Section 2.2.3) in order to verify the performance of their algorithms. The train/validation stage lasted approximately 80 days, after which submissions for the train/validation data were no longer accepted. The test stage could then begin with the release of the test data and participants were given 10 days to submit test data results. On both the training/validation and test stages participants were limited to one submission per 24-hour period to prevent overfitting to the test set.

Given the existence of other public datasets relevant for the tasks of the LNDb Challenge, the use of external data was allowed at both stages of the challenge, with the condition that the use of external data be reported upon publication.

2.2. Dataset Description

2.2.1. Data Acquisition

The LNDb Challenge was based in the homonymous LNDb dataset, publicly released specifically for this challenge. The LNDb dataset contains 294 CT scans collected retrospectively at the Centro Hospitalar e Universitário de São João (CHUSJ) in Porto, Portugal between 2016 and 2018. All data was acquired under approval from the CHUSJ Ethical Commitee and was anonymised prior to any analysis to remove personal information except for patient birth year and gender. Further details on patient selection and data acquisition can be consulted on the database description paper (Pedrosa et al., 2019).

2.2.2. Data Annotation

Each CT scan was read by between one and three radiologists at CHUSJ to identify pulmonary nodules and other suspicious lesions. A total of 5 radiologists with at least 4 years of experience participated in the annotation process. Annotations were performed in single blinded fashion, i.e. a radiologist would read the scan once and no consensus or review between the radiologists was performed. The instructions for manual annotation were adapted from LIDC-IDRI so that each radiologist identified the following findings:

- nodule ≥3mm: any lesion considered to be a nodule by the radiologist with greatest in-plane dimension larger or equal to 3mm;
- nodule <3mm: any lesion considered to be a nodule by the radiologist with greatest in-plane dimension smaller than 3mm;
- non-nodule: any pulmonary lesion considered not to be a nodule by the radiologist, but that contains features which could make it identifiable as a nodule.

The annotation process varied for the different categories. Nodules ≥3mm were segmented and subjectively characterized according to LIDC-IDRI (ratings on subtlety, internal structure, calcification, sphericity, margin, lobulation, spiculation, texture and likelihood of malignancy). For a complete description of these characteristics the reader is referred to McNitt-Gray et al. (2007). For nodules <3mm the nodule centroid was marked and subjective assessment of the nodule's characteristics was performed. For non-nodules, only the lesion centroid was marked. Given that different radiologists may have read the same CT and no consensus review was performed, variability in radiologist annotations is expected.

[1] `https://lndb.grand-challenge.org/`

2.2.3. Data Curation

CT scans were made available on MetaImage (.mhd/.raw) format. Nodule and non-nodule annotations were made available in two different .csv files. The first file listed all findings as annotated by all radiologists. For each finding the following information was provided: the ID of the finding, the xyz coordinates of the finding in world coordinates, whether it is a nodule or a non-nodule, its volume according to the segmentation and the nodule texture rating (1-5). For non-nodules, the texture given was 0. The second file listed all unique findings after merging findings annotated by different radiologists. Findings annotated by different radiologists in the same CT scan were considered to be a unique finding if the Euclidean distance between their centroids was smaller or equal than the maximum equivalent diameter (the diameter of a sphere with a volume equal to the nodule volume) of the two findings. For findings of equivalent diameter smaller than 3mm, an equivalent diameter of 3mm was considered. Findings marked as a nodule by a single radiologist were considered to be a nodule independent of other radiologist annotations. For each unique finding the same information as on the first .csv file was given, except that the volume and texture were the average of the volume and texture among the radiolists that annotated each finding. Additionally, the number of radiologists that annotated each finding was also given.

Nodule segmentation masks were also given on MetaImage (.mhd/.raw) as a 3D array of the CT's size where each finding was identified by the ID given in the .csv file with all annotations. For each CT, a separate segmentation mask per radiologist was given so that each mask contained the segmentations for all nodules on that CT scan according to that radiologist.

Finally, ground truth Fleischner classification was computed for each CT scan and made available on a third .csv file. The original Fleischner classification was recast into the following four classes:

0. No routine follow-up required or optional CT at 12 months according to patient risk;
1. CT at 6-12 months required;
2. CT at 3-6 months required;
3. CT, PET/CT or tissue sampling at 3 months required.

Table 1: Fleischner classification rules used.

Single Nodule	Volume $<100mm^3$	Volume $100–250mm^3$	Volume $\geq 250mm^3$
GGO	0	1	1
Part-solid	0	2	2
Solid	0	1	3
Multiple Nodules	Volume $<100mm^3$	Volume $100–250mm^3$	Volume $\geq 250mm^3$
GGO/Part-solid	2	2	2
Solid	0	2	2
Mixed	Classify for GGO/part-solid and solid nodules independently and attribute highest class		

The Fleischner score was computed directly from the radiologist nodule annotations according to a set of rules (MacMahon et al., 2017) taking into account the number of nodules (single or multiple), their volume ($< 100mm^3$, $100–250mm^3$ and $\geq 250mm^3$) and texture (solid, part solid and GGO) as shown in Table 1. Note that while the Fleischner guidelines also take into account patient risk factors (such as age, sec, race, family history, smoking history and others), this information was not available and was thus not taken into account. Before computing the Fleischner score for each CT, the annotations of each radiologist were merged into a list of unique findings as described above. For each unique finding, nodule volume was considered to be the average volume of the segmentation of each radiologist for a given nodule and nodule texture was recast from the five classes in the LNDb annotation (1-GGO, 2-intermediate, 3-part solid, 4-intermediate, 5-solid) into the three classes of the Fleischner guidelines by considering GGO as 1-2, part solid as 3 and solid as 4-5. If multiple radiologists identified the nodule, the average texture was computed and the three classes of the Fleischner guidelines were computed by considering GGO as $< 7/3$, part solid as $7/3 - 11/3$ and solid as $> 11/3$.

Finally, the full LNDb dataset was divided into five folds. The first four were used for train/validation and the fifth was reserved for testing. A balanced distribution of Fleischner classes and of the number of radiologists that annotated each CT scan were maintained across all folds, except for the test subset for

which only CTs annotated by at least two radiologists were selected. A list of the CTs belonging to each fold was provided to the participants, allowing participants to perform four-fold cross-validation during the train/validation stage for a robust assessment of performance. The data and annotations of the four train/validation folds were released at the beginning of the train/validation stage, whereas only the data (and not the annotations) of the test fold was released at the beginning of the test stage.

2.3. Challenge Tasks

2.3.1. Main Challenge - Fleischner Classification

The main challenge was the automatic classification of CT scans according to the 2017 Fleischner society pulmonary nodule guidelines for patient follow-up recommendation. For a given CT scan, participants should predict the final Fleischner classification. Only the predicted Fleischner class was taken into account for evaluation and the detection of nodules, their segmentation and texture characterization was not taken into account, allowing for the use of end-to-end solutions.

Participants were asked to submit, for each CT scan, a probability value (0-1) for each Fleischner class. For evaluation, the class with maximum probability was treated as the predicted Fleischner class and if two classes had equal and maximum probability, the class with higher index was treated as the predicted Fleischner class. The submitted Fleischner predictions were compared to the ground truth and the agreement was computed according to Fleiss-Cohen weighted Cohen's kappa Spitzer et al. (1967)

$$\kappa_w = \frac{\sum_i^k \sum_j^k w_{ij} p_{ij} - \sum_i^k \sum_j^k w_{ij} p_{i*} p_{*j}}{1 - \sum_i^k \sum_j^k w_{ij} p_{i*} p_{*j}} \quad (1)$$

where p_{ij} is the proportion of cases with ground truth class i and rated by the participant as class j. $*$ is a wildcard so that p_{*j} is the proportion of cases rated by the participant as class j. w_{ij} is the weight for class combination ij according to

$$w_{ij} = \frac{(C_i - C_j)^2}{(C_1 - C_k)^2} \quad (2)$$

for a rating with k classes (C_1,C_2,...,C_k).

2.3.2. Challenge A - Nodule Detection

Challenge A was the automatic detection of pulmonary nodules in CT scans. All nodules, independent of their size and characteristics (including nodules <3mm) should be detected. The merged list of unique nodule findings was used as ground truth, obtained as described in 2.2.3.

Participants were asked to submit, for each CT scan, a list of all nodule candidates with the corresponding xyz coordinates and the predicted probability of the candidate being a nodule (0-1). For evaluation, the submitted nodule candidates were compared to the ground truth annotations. A candidate was considered a true positive if the Euclidean distance between the predicted centroid and a ground truth nodule centroid was smaller or equal than the maximum equivalent diameter of the ground truth nodule. For nodules of equivalent diameter smaller than 3mm, an equivalent diameter of 3mm was considered. True nodules for which no nodule candidate followed the above rule were considered false negatives. A candidate was considered an FP if there was no finding following the above rule. A candidate that matched a non-nodule was also considered an FP.

Similarly to previous challenges on lung nodule detection (Van Ginneken et al., 2010; Setio et al., 2017), evaluation was performed on the free receiver operating characteristic (FROC) curve. The mean sensitivity $\bar{s}$ was computed at 7 predefined false positive rates: 1/8, 1/4, 1/2, 1, 2, 4, and 8 FPs per scan:

$$\bar{s} = \frac{1}{7} \sum_{i \in FP} s(i), \;\; FP = \{1/8, 1/4, 1/2, 1, 2, 4, 8\}, \quad (3)$$

where $s(i)$ is the sensitivity for FP rate i.

To account for observer variability, average sensitivity was computed at different agreement levels. Two different FROC curves are computed considering: 1) all nodules (agreement level 1); 2) nodules marked by at least two radiologists (agreement level 2). In this way, the more consensual nodules, i.e. those marked by a higher number of radiologists, have a larger weight on the final score (as they appear on both agreement levels). The final ranking of the different methods was obtained according to score s_A computed as the average of the FROC

average sensitivity at both agreement levels:

$$s_A = \frac{\bar{s}_1}{2} + \frac{\bar{s}_2}{2} \tag{4}$$

where $\bar{s}_k$ is the mean sensitivity at the predefined FP rates for agreement level k.

2.3.3. Challenge B - Nodule Segmentation

Challenge B was the automatic segmentation of pulmonary nodules ≥3mm in CT scans.

Participants were asked to submit the segmentation of every unique pulmonary nodule ≥3mm as computed in Section 2.2.3. However, to prevent participants from using the list of ground truth nodules for the Main Challenge and Challenge A, which would invalidate both these challenges, this list was mixed with a high number of FPs. The FPs were obtained from the automatic nodule detection framework proposed in Aresta et al. (2018) and randomly selecting nodule candidates with a low predicted probability of being a true nodule until a total of 50 centroids per CT were obtained. Nodule segmentations were submitted as $80 \times 80 \times 80$ cubes with voxelsize 0.6375mm centered on the nodule centroid and the biggest connected object was treated as the predicted nodule segmentation.

For evaluation of the accuracy of the segmentation, three segmentation performance measures were considered:

- Modified Jaccard index (J^*) computed as a measure of overlap between the predicted segmentation volume (V) and the reference segmentation volume (V_r):

$$J^* = 1 - \frac{V \cap V_r}{V \cup V_r}; \tag{5}$$

- Mean Average Distance (MAD) between the predicted surface (S) and the reference surface (S_r):

$$MAD = \frac{1}{2}(d_{mean}(S, S_r) + d_{mean}(S_r, S)), \tag{6}$$

  where $d_{mean}(S_1, S_2)$ is the mean of distances between every surface voxel in S_1 and the closest surface voxel in S_2;

- Hausdorff Distance (HD) between the predicted surface (S) and the reference surface (S_r):

$$HD = max(d_{max}(S, S_r) + d_{max}(S_r, S)), \tag{7}$$

  where $d_{max}(S_1, S_2)$ is the maximum of distances between every surface voxel in S_1 and the closest surface voxel in S_2;

Furthermore, to measure the degree of accuracy of the segmentation for extraction of clinical indices, three volume performance measures were computed comparing the predicted and reference volumes:

- Modified Pearson correlation coefficient $r^* = 1 - r$ where r is the Pearson correlation coefficient between the predicted and reference volumes;
- Bias (b) computed as the mean absolute difference of the predicted and reference volumes;
- Standard deviation (σ) of the difference of the predicted and reference volumes.

Given that each nodule can be annotated by multiple radiologists, and thus have multiple segmentation ground truths, J^*, MAD and HD were computed in reference to the segmentation of each radiologist and then averaged per nodule. In this way, a nodule annotated by several radiologists has the same weight for the final score as a nodule annotated by a single radiologist. However, r^*, b and σ were computed in comparison to the average volume obtained from the segmentations of all radiologists.

The final ranking was obtained according to score s_B calculated as the average of all the six measures normalized according to the maximum among all participants (indicated by the symbol $'$) so that each individual measure takes a value between 0 (worst case among all participants) and 1 (perfect fit between the reference and the predicted segmentation):

$$s_B = \frac{J^{*\prime} + MAD' + HD' + r^{*\prime} + b' + \sigma'}{6}. \tag{8}$$

2.3.4. Challenge C - Nodule Texture Characterization

Challenge C was the automatic characterization of texture of all pulmonary nodules (both <3mm and ≥3mm). Three texture classes were considered following the classification in the Fleischner guidelines : 0) Ground glass opacities (GGO), 1) Part solid nodules (PSN), 2) Solid nodules (SN). Ground truth nodule texture was computed as outlined in Section 2.2.3 for Fleischner classification.

Table 2: Summary of the participants with successful submissions on the test stage and corresponding challenges participated.

Participant (*alias*)	Challenges			
	Main	A	B	C
Atwal and Phoulady (*atwalg*)			✓	✓
Chen et al. (*LINK*)				✓
Galdran and Bouchachia (*agaldran*)			✓	✓
Kaluva et al. (*nightfury*)	✓	✓	✓	✓
Katz et al. (*IRC*)		✓		
Kim (*ildoo*)	✓			
Rassadin (*alexander.rassadin*)			✓	✓
Sun et al. (*Mediclouds*)			✓	✓
??? (*eddie*)*			✓	
??? (*Look*)**			✓	✓
??? (*medi-perk*)			✓	

* Did not complete the train/validation stage submission for Challenge B.

** Did not complete the train/validation stage submission for Challenge C.

Participants were asked to submit, for each nodule, the probability of belonging to each of the three texture classes. For the test set, the same methodology as in Challenge B was used in order not to invalidate the Main Challenge and Challenge A, by adding to the list of ground truth nodules a high number of FPs. For evaluation, the class with maximum probability was treated as the predicted texture class and if two classes had equal and maximum probability, the class with higher index was treated as the predicted class. The submitted texture predictions were compared to the ground truth and agreement was computed according to Fleiss-Cohen weighted Cohen's kappa (Spitzer et al., 1967) described on equation 1.

3. Challenge Participations

The LNDb Challenge had a total of 847 participations. A total of 197 individual submissions were made, with 25 participants making a valid submission on the train/validation stage and 20 participants making a submission for the test stage. A total of 11 participants successfully made a submission for the test stage, 10 of which also made a submission for the train/validation stage, as shown in Table 2. All 11 participants were contacted by email for participation in this manuscript, of which four (*eddie*, *Look* and *medi-perk*) could not be contacted, which did not allow the authors for a description of those participants' methods.

Tables 3, 4, 5 and 6 summarize the approaches used by each participant in each of the challenges. A more detailed description of each approach is given below.

3.1. Atwal and Phoulady (atwalg)

Atwal and Phoulady (2020) participated on challenges B and C. All CT images were resampled such that each voxel had an isotropic size of 0.6375mm^3. The values of each resampled image were clipped to a Hounsfield Units (HU) range of [-1000, 500], zero-centered by subtracting the mean value, and min-max normalized between 0 and 255. Patches of size 51mm^3 centered on the nodule centroids were then extracted and used as input for two CNNs. The segmentation network followed the U-net architecture introduced by Ronneberger et al. (2015) but used 3D layers instead of 2D and used half the number of channels for each convolutional layer to compensate for the extra dimension. The network predicted a probability for each value in the input to produce a mask. The texture characterization network was based on a VGG architecture (Simonyan and Zisserman, 2014) and used four blocks of 3D convolutional layers followed by two fully-connected layers to predict the probability of each texture class.

3.2. Chen et al. (LINK)

Chen et al. participated on challenge C. All CT images were preprocessed to enlarge the details of nodule texture. In particular, each 80×80×80 nodule cube was split into a high number of $30\times30\times3$ slices along the horizontal, coronal and sagittal plane, and further interpolated into $224\times224\times3$. In order to enrich the diversity of texture representations and expand the training dataset, 2.5D representations (through concatenation of horizontal, coronal and sagittal slices) were used for classification. A deep pre-trained 2D network (SE-ResNet101 (He et al., 2016)) was then applied as the classifier, reducing the complexity of feature parameters compared to a 3D CNN (Dey et al., 2018), without losing 3D context. Finally, an ensemble of four models from the cross-validations was used to enhance the robustness of predictions.

3.3. Galdran and Bouchachia (agaldran)

Galdran and Bouchachia (2020) participated on challenges B and C. For texture categorization, nodules were sampled from CT scans and three orthogonal planes passing through nodule centroids are extracted during training. A ResNet50 (He et al., 2016) network was trained for texture classification and Gaussian Label Smoothing (Galdran et al., 2020), a regularization scheme based on a custom manipulation of manual labels that better optimizes κ_w by penalizing predictions further away from the correct class, was used during training. For the nodule segmentation task, three-dimensional nodules are employed. A modified 3-D U-Net architecture was constructed by adding residual connections inside each of its blocks, and also add convolutional operations to skip connections from the downsampling path to the upsampling path. For texture classification, the Cross-Entropy loss applied on smoothed labels was backpropagated, whereas for nodule segmentation, a 3D dice loss was employed. In both cases, training was performed for 500 epochs and Test-Time Augmentation (Wang et al., 2019) was applied to generate predictions.

3.4. Kaluva et al. (nightfury)

Kaluva et al. (2020) participated on all challenges. Kaluva et al. proposed a 5-stage deep learning approach, including lung segmentation, nodule detection, nodule texture classification, nodule segmentation and follow-up recommendation. A 3D U-net (Ronneberger et al., 2015) was first used to segment lungs from the CT scan. Within the segmented lung region, 3D cubic patches of size 132 were extracted and passed to a 3D FasterRCNN (Ren et al., 2015) to predict nodule candidates which were then fed to a 3D WideResNet (Zagoruyko and Komodakis, 2016) to classify them as nodule / non-nodule. 3D patches around all detected nodules were extracted to (a) classify texture using a WideResnet and (b) segment the nodule using a U-net. The predicted segmentation nodule volume and nodule texture were then used to predict the follow-up in accordance with the Fleischner guidelines. All the models were trained using LIDC-IDRI (Armato et al., 2011), NLST (The National Lung Screening Trial Research Team, 2011) and LNDb.

3.5. Katz et al. (IRC)

Katz et al. (2020) participated on challenge A. Katz et al. proposed an ensemble learning pipeline based on 3D SE-ResNet18 (He et al., 2016) and DPN68 (Chen et al., 2017). CTs were resampled to isotropic 1mm voxels and clipped to [−1200,600] HU. Lung segmentation was then performed using a 3D U-Net (Çiçek et al., 2016) trained on the LOLA11 dataset (Van Rikxoort and Van Ginneken, 2011). The nodule detection framework is composed of two independent detectors, a SE-ResNet18 and a DPN68. Candidates from each of the detectors were merged through empirically defined thresholds on the predicted probability of each nodule and were then passed to an FP reduction module. The FP reduction module is a custom built CNN similar to a pyramidal CNN which takes as input 3D volumes of three different sizes (16, 24 and 48mm) and the final nodule/non-nodule classification is given via majority vote. Nodule detection and FP reduction were trained on the LUNA16 and Kaggle Data Science Bowl[2] datasets and 3D data augmentation was performed during training, namely flipping, scaling, rotation, and random HU perturbation transformations.

3.6. Kim (ildoo)

Kim participated on the main challenge. Kim performed lung nodule segmentation and texture characterization for each 3D lung CT scan, using a 3D U-Net variant. Labels from several radiologists were averaged and preprocessed into an agreed label. Since the intensity of CT images is absolute, z-score normalized input with clipped values out of 99 percentile to zero was used. The network architecture and training method follows the method proposed by Isensee et al. (2018). nnU-Net has heuristic logics to determine hyperparameters, such as input shape, to work well with multiple 3D biomedical image segmentation. The basic hyperparameters of nnU-Net were tuned and the setting that produced the best performance consistently in 5-fold cross validation was used on test. The best performance was obtained by applying an ensemble technique on rotation/flip test-time augmentations and 5-fold models. As a result of predict-

[2] `https://www.kaggle.com/c/data-science-bowl-2017`

ing lung nodule segmentation and texture, the follow-up recommendation was then calculated according to the Fleischner guidelines.

3.7. Rassadin (alexander.rassadin)

Rassadin (2020) participated on challenges B and C. Rassadin proposed a joint nodule segmentation and texture classification neural network, exploiting the idea of so-called multi-task learning. The network is a deep residual U-Net (Ronneberger et al., 2015) with batch normalization (Ioffe and Szegedy, 2015) replaced by a group normalization (Wu and He, 2018) and rectified linear unit activations (ReLu) (Glorot et al., 2011) replaced by exponential linear unit activations (ELU) (Clevert et al., 2015). A fully connected network then receives the output of the encoder section of the U-Net to predict nodule texture. Training of the segmentation and texture classification branch was then performed simultaneously.

3.8. Sun et al. (Mediclouds)

Sun et al. (2020) participated on challenges B and C. First, Models Genesis (Zhou et al., 2019), a self-supervised learning method, was applied to obtain a pre-trained model which can be used for both image classification and segmentation. A 3D U-net (Çiçek et al., 2016) was adopted as the network architecture for pre-training, nodule segmentation and classification. In the pretraining procedure, cubes were extracted at random positions within the CT scans, to which 3 kinds of noises were randomly added. The images with added noise and the original images were used as input and target data respectively. The network was pretrained using mean squared error as loss function. In the segmentation task, the pretrained 3D U-Net was used for nodule segmentation training and BCE and Dice as loss functions. In the classification task, there were two path of inputs. The first was the U-Net encoder, after which global max pooling was applied to the final feature map. The second was based on the segmentation result by extracting features from each nodule cube, namely 50 values of the histogram of the segmented nodule in a [-1350; 150] HU window, gray value variance and mean and nodule volume. The extracted features were concatenated with the output of global max pooling layer, followed by an fully connected and softmax layer. Categorical cross entropy was used as loss function during training. In addition, 3D data augmentation (flipping, rotation shifting and zoom) in all training stages was used to prevent over fitting. Adam was used as optimizer in all experiments.

Table 3: Summary of the approaches submitted for the main challenge. HU: HU window used; Voxelsize: voxelsize of input data used; Lung segmentation: whether lung segmentation was applied as FP reduction mechanism; Approach: main methods used; Input size: data input size in voxels used (axial×coronal×sagittal); Data: datasets used for training/validation.

Participant	Kaluva et al.	Kim
Data	LIDC-IDRI NLST; private	LNDb
HU	$[-1200, 400]$	[−1200,Inf]
Voxelsize	1×1×1mm	1×1×1mm
Fleischner Classification		
Approach	Rule based	Rule based
Nodule Detection	cf. Table 4	nnU-Net
Nodule Segmentation	cf. Table 5	nnU-Net
Texture Characterization	cf. Table 6	nnU-Net

Table 4: Summary of the approaches submitted for challenge A. Data: datasets used for training/validation; HU: HU window used; Voxelsize: voxelsize of input data used (only one dimension is given for isotropic voxels); Lung segmentation: whether lung segmentation was applied as FP reduction mechanism; Approach: main methods used; Input size: data input size in voxels used (axial×coronal×sagittal).

Participant	Kaluva et al.	Katz et al.
Data	LIDC-IDRI; NLST; private	LIDC-IDRI; Kaggle
HU	$[-1200, 400]$	$[-1200, 600]$
Voxelsize	1mm	1mm
Lung Segmentation	✓	✓
Nodule Detection		
Approach	Faster R-CNN	DPN68+SE-ResNet18
Input size	128×128×128	Unknown
FP Reduction		
Approach	3D WRN	Custom CNN
Input size	64×64×64	16×48×48

4. Experiments

To complement the performance evaluation performed on each challenge as described in Section 2.3, interobserver variability between the five radiologists in all four challenges was computed. For the main challenge and challenge C, the κ_w between each radiologist and the ground truth was computed. Note that for each radiologist, only scans/nodules annotated by that radiologist and at least one other radiologist were consid-

Table 5: Summary of the approaches submitted for challenge B. Data: datasets used for training/validation; HU: HU window used; Voxelsize: voxelsize of input data used (only one dimension is given for isotropic voxels); Approach: main methods used; Pretraining: whether pretraining or a pretrained network was used and which; Joint: whether a joint training or multi-task strategy was applied. Input size: data input size in voxels used (axial×coronal×sagittal).

Participant	Data	HU	Voxelsize	Approach	Pretraining	Joint	Input size
Atwal and Phoulady	LNDb	[−1000, 500]	0.6375mm	3D U-Net	✗	✗	80×80×80
Galdran and Bouchachia	LNDb	[−1000, 1000]	0.6375mm	3D U-Net	✗	✗	80×80×80
Kaluva et al.	LIDC-IDRI	[−1200, 400]; [−125, 225]	1mm	3D U-Net	✗	✗	132×132×132
Rassadin	LNDb	[−Inf,+Inf]	0.6375mm	3D U-Net	✗	Texture	80×80×80
Sun et al.	LIDC-IDRI; TIANCHI; LNDb	[−1150, 350]	0.6375mm	3D U-Net	Models Genesis	Texture	80×80×80

Table 6: Summary of the approaches submitted for challenge C. Data: datasets used for training/validation; HU: HU window used; Voxelsize: voxelsize of input data used (only one dimension is given for isotropic voxels); Approach: main methods used; Pretraining: whether pretraining or a pretrained network was used and which; Joint: whether a joint training or multi-task strategy was applied. Input size: data input size in voxels used (axial×coronal×sagittal).

Participant	Data	HU	Voxelsize	Approach	Pretraining	Joint	Input size
Atwal and Phoulady	LNDb	[−1000, 500]	0.6375mm	Custom CNN	✗	✗	80×80×80
Chen et al.	LIDC-IDRI; LNDb	[−1000, 400]	0.6375mm	SE-ResNet101	ImageNet	✗	3 orthogonal 30×30
Galdran and Bouchachia	LNDb	[−1000, 1000]	0.6375mm	ResNet50	✗	✗	3 orthogonal 64×64
Kaluva et al.	LIDC-IDRI; NLST; LNDb	[−1200, 400]	1mm	3D WRN	✗	✗	64×64×64
Rassadin	LNDb	[−Inf,+Inf]	0.6375mm	3D U-Net encoder+FCN	✗	Segmentation	80×80×80
Sun et al.	LIDC-IDRI; TIANCHI; LNDb	[−1150, 350]; [−1350, 150]	0.6375mm	3D U-Net encoder +gray values+FCN	Models Genesis	Segmentation	80×80×80

ered. For challenge A, the sensitivity and FP/scan of each radiologist was computed by considering the other four radiologists as ground truth and applying the same rules as described in 2.3. For challenge B, interobserver variability was computed for every nodule annotated by more than one radiologist by computing the average J^*, MAD, HD and absolute volume difference of all possible combinations of radiologist segmentations of that nodule. Note that r^* and σ cannot be computed by averaging across all radiologists and are thus not reported.

Furthermore, where applicable, the statistical significance of the difference between the different participants and the interobserver variability was tested. For the main challenge and challenge C, an adaptation of McNemar's test (Edwards, 1948) was used to assess the difference in classification accuracy according to:

$$\chi^2 = \frac{(|n_{01} - n_{10}| - 1)^2}{n_{01} + n_{10}}, \tag{9}$$

where n_{01} is the number of misclassified samples by method 1 but not by method 0 and n_{10} viceversa. McNemar tests were computed for $p<0.05$ ($\chi^2>3.841$) and $p<0.01$ ($\chi^2>6.635$). For challenge B, paired t-test was performed for J^*, MAD and HD for $p<0.05$ and $p<0.01$.

5. Challenge Results

5.1. Main Challenge - Fleischner Classification

Table 7 shows the ranking of the participants of the main challenge on the train/validation and the test stages. Kim obtained the best results on both stages, though the difference to Kaluva et al. was only statistically significant at the

Table 7: Fleischner classification results on the train/validation and test stages. Best results in each stage shown in bold. Interobserver variability reported as mean ± standard deviation. $*$ and $**$ indicate a statistical significant difference to the best participant within each stage according to McNemar test at $p<0.05$ and $p<0.01$ respectively. ^{N} indicates a statistical significant difference to N out of the 5 radiologists according to McNemar test at $p<0.05$.

Participant	Train/Validation κ_w	Test κ_w
Kaluva et al.	$0.532^{**,4}$	0.464^{1}
Kim	$\mathbf{0.603}^{3}$	**0.580**
Interobserver	0.743±0.152	0.604±0.240

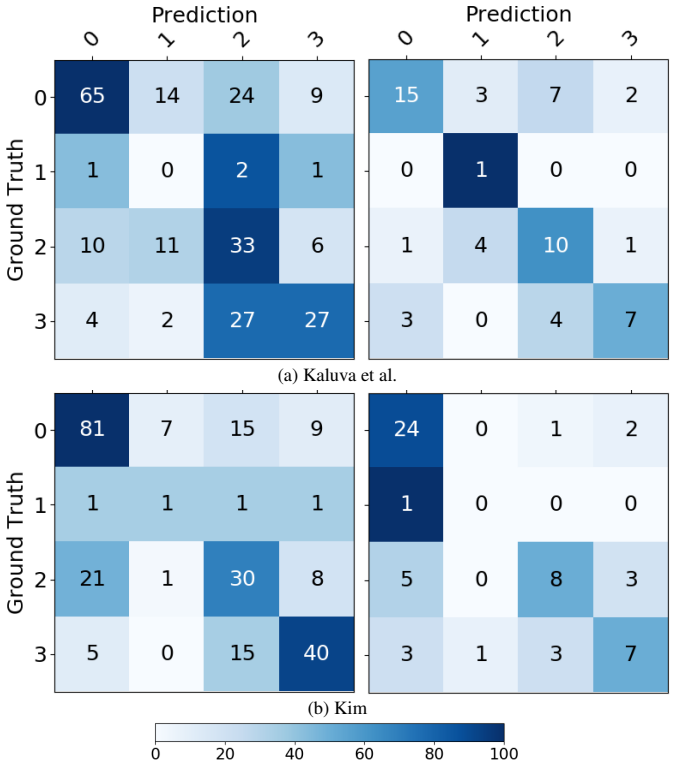


Fig. 1: Confusion matrix of Fleischner classes on train/validation (left) and test (right) stages for each participant. Color intensity corresponds to the percentage of CTs within each class.

train/validation stage. Figure 1 shows the confusion matrix on train/validation and test stages for the four Fleischner classes.

5.2. Challenge A - Nodule Detection

Table 8 shows the classification of the participants of challenge A on the train/validation and the test stages. Figure 2 shows the FROCs obtained at train/validation and test stages. While Katz et al. obtained a better sensitivity during train/validation, the performance was somewhat degraded during test. The opposite trend was observed for Kaluva et al., leading to a better performance on the test set. Figure 3 shows examples of nodule candidates from the test stage evaluation

Table 8: Nodule detection results on the train/validation and test stages. Best results in each stage and measure shown in bold.

Participant	Train/Validation s_A	$\bar{s}_1$	$\bar{s}_2$
Kaluva et al.	0.348	0.268	0.4279
Katz et al.	**0.473**	**0.364**	**0.582**
Participant	**Test** s_A	$\bar{s}_1$	$\bar{s}_2$
Kaluva et al.	**0.533**	**0.396**	**0.671**
Katz et al.	0.442	0.356	0.528

extracted at different FP/scan levels for each participant. It can be seen that vessels and other structures such as atelectasis are those most confused with true nodules but bone structures can also be observed at higher FP/scan levels for Katz et al.'s approach.

5.3. Challenge B - Nodule Segmentation

Table 9 shows the classification of the participants of challenge B on the train/validation and the test stages. Figure 4 shows the Bland-Altman plots for nodule volume obtained on the test stage. While *Look* obtained the best results in terms of segmentation performance measures on the train/validation stage, these results were not replicated on the test stage where a significant overestimation of nodule volume can be observed. The two best methods on both train/validation and test stages were those by Rassadin and Sun et al.. In spite of neither of these methods having the best results in terms of segmentation performance measures, they have a smaller r^*, bias and standard deviation to ground truth nodule volumes. Figure 5 shows examples of nodule segmentations from the test stage evaluation extracted at different agreement levels for each participant.

5.4. Challenge C - Nodule Texture Characterization

Table 10 shows the evaluation results of challenge C on the train/validation and the test stages. Figure 6 shows the receiver operating characteristic curves (ROC) and confusion matrixes obtained on the test stage for each participant. While Atwal and Phoulady obtained the best results on the train/validation stage, performance was significantly degraded on the test stage as seen on both the ROC and confusion matrix. The best results

Table 9: Nodule segmentation results on the train/validation and test stages. Best results in each stage and measure shown in bold. Data is mean ± standard deviation where applicable. $*$ and $**$ indicate a statistical significant difference to the best participant within each stage and measure according to paired t-test at $p<0.05$ and $p<0.01$ respectively. † and ‡ indicate a statistical significant difference to the interobserver variability according to paired t-test at $p<0.05$ and $p<0.01$ respectively. Note that interobserver variability and correponding t-test can only be calculated for nodules annotations by at least two radiologists.

Participant	Train/Validation						
	s_B	J^*	$MAD\,(mm)$	$HD\,(mm)$	r^*	b	σ
Atwal and Phoulady	0.599	0.601±0.175**,‡	1.337±3.328**,‡	4.130±4.990**,‡	0.125	220.57	682.98
Galdran and Bouchachia	0.714	0.432±0.150**,‡	0.458±0.573**	2.236±2.260**	0.122	125.46	706.67
Kaluva et al.	0.372	0.789±0.243**,‡	6.045±7.815**,‡	9.387±9.308**,‡	0.145	204.95	683.55
Rassadin	0.742	0.489±0.170**	0.567±0.990**	2.482±2.246**	**0.078**	103.32	**486.87**
Sun et al.	**0.771**	0.433±0.126**,‡	0.389±0.422**,‡	2.049±1.597**,‡	0.079	**75.54**	507.37
eddie	—	—	—	—	—	—	—
Look	0.740	**0.369±0.117**‡	**0.348±0.446**‡	**1.899±1.876**†	0.112	111.46	719.58
medi-perk	0.191	0.839±0.168**,‡	4.876±5.613**,‡	14.461±7.110**,‡	0.287	426.84	876.66
Interobserver		0.467±0.185	0.447±0.233	2.162±1.181	—	88.29	—

Participant	Test						
	s_B	J^*	$MAD\,(mm)$	$HD\,(mm)$	r^*	b	σ
Atwal and Phoulady	0.578	0.766±0.209**,‡	0.987±0.690**,‡	3.277±1.586**,‡	0.118	83.04	119.86
Galdran and Bouchachia	0.725	0.445±0.144‡	0.412±0.258‡	2.062±1.502‡	0.145	41.43	129.47
Kaluva et al.	0.481	0.597±0.236**	2.322±5.931**	4.406±6.984**	0.169	77.37	239.41
Rassadin	**0.754**	0.478±0.478**,†	0.420±0.215	2.028±1.229‡	**0.055**	44.28	**86.32**
Sun et al.	0.743	0.468±0.136**	0.469±0.798	2.137±1.514	0.081	**40.70**	98.74
eddie	0.715	**0.433±0.139**‡	**0.397±0.279**	**1.983±1.463**	0.175	44.63	141.49
Look	0.044	0.562±0.246**	3.561±7.087**,†	7.462±10.715**,‡	0.612	209.13	535.62
medi-perk	0.434	0.614±0.249**	3.359±8.379**	5.729±9.318**	0.175	70.32	140.83
Interobserver		0.494±0.200	0.470±0.239	2.163±1.101	—	53.25	—

Table 10: Nodule texture characterization results on the train/validation and test stages. Best results in each stage shown in bold. Interobserver variability reported as mean ± standard deviation. $*$ and $**$ indicate a statistical significant difference to the best participant within each stage according to McNemar test at $p<0.05$ and $p<0.01$ respectively. ^{N} indicates a statistical significant difference to N out of the 5 radiologists according to McNemar test at $p<0.05$.

Participant	Train/Validation	Test
	κ_w	κ_w
Atwal and Phoulady	**0.904**^{1}	-0.008^{**,5}
Chen et al.	0.427^{**,3}	**0.733**
Galdran and Bouchachia	0.568^{**,3}	0.613
Kaluva et al.	0.369^{**,3}	0.3407^{1}
Rassadin	0^{**,3}	0.0279^{**,5}
Sun et al.	0.679^{**,3}	0.686^{1}
Look	—	0.695^{1}
Interobserver	0.738±0.135	0.870±0.106

on the test stage were obtained by Chen et al., though similarly good performance was obtained by Sun et al. and *Look*. Figure 7 shows examples of nodule texture characterization from the test stage evaluation extracted at different predicted ground truth texture class predicted probability for each participant.

6. Discussion

6.1. Main Challenge - Fleischner Classification

Comparing the results of Kim and Kaluva et al. on the main challenge, Figure 1 shows that Kaluva et al.'s method has significantly more difficulty in identifying class 0 (no follow-up or optional CT at 12 months), often misclassifiying those cases as classes 1-3. Since both methods are rule based, i.e. they rely on nodule detection, segmentation and texture characterization followed by a direct application of the Fleischner guidelines, the differences in performance between the two methods are directly related to the nodule detection, segmentation and/or texture characterization. However, because Kim did not participate in challenges A, B and C, a direct comparison of each separate challenge cannot be made. Nevertheless, it can be seen that Kaluva et al. had a good performance in nodule detection but below average performance in both nodule segmentation and texture characterization. Nodule segmentation in particular may have played a significant role given that, as shown in

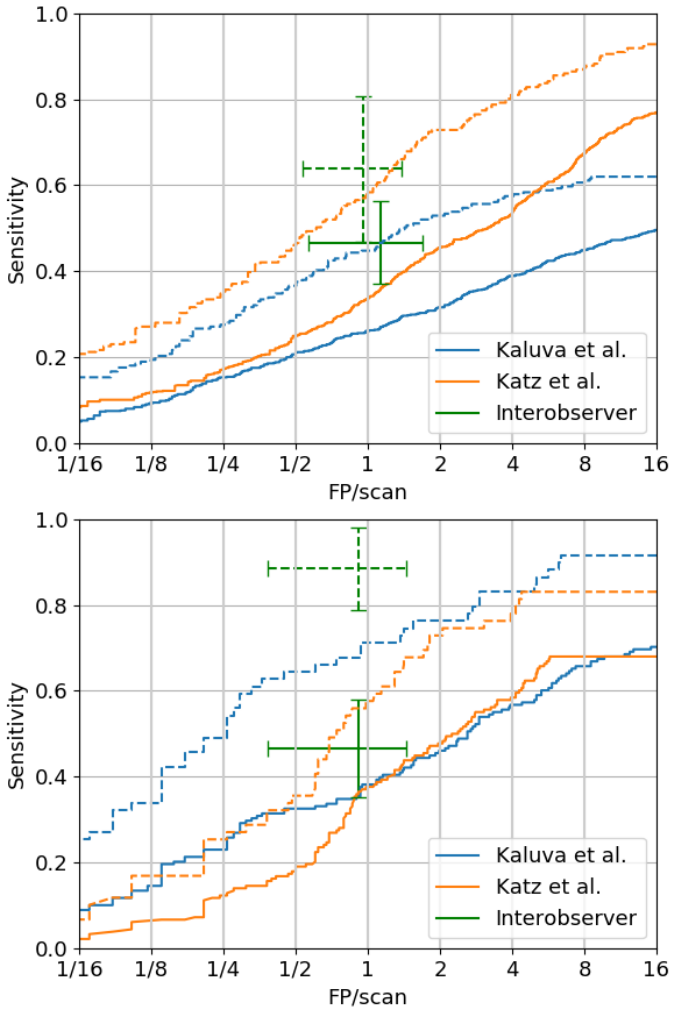


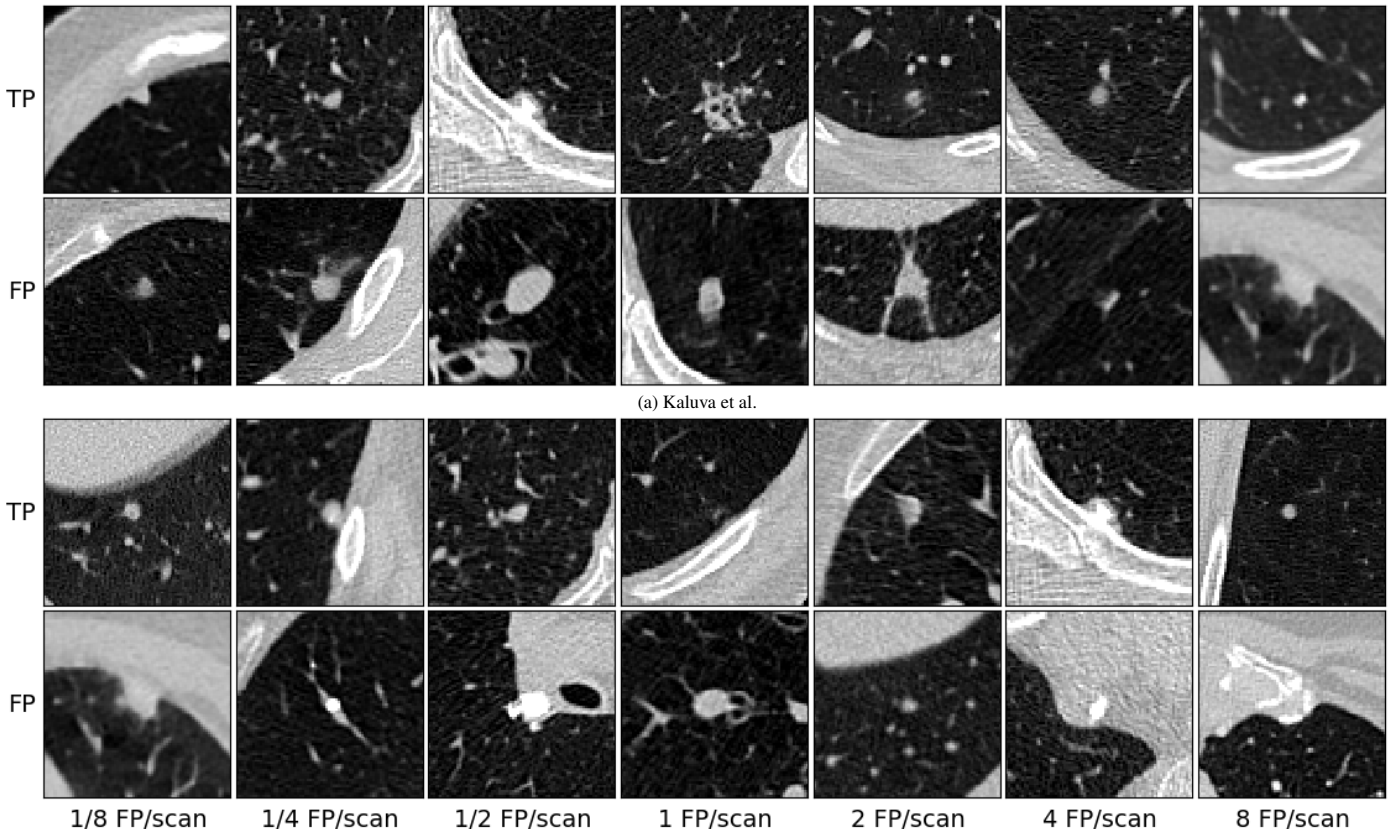
Fig. 2: Nodule detection FROCs on train/validation (left) and test (right) stages. Full and dashed lines correspond to agreement levels 1 and 2 respectively. Interobserver bars correspond to mean ± standard deviation sensitivity and FP/scan for the five radiologists.

Figure 4, Kaluva et al.'s method tended to overestimate nodule volume in smaller nodules, which hinders the correct classification of Fleischner class 0 according to the classification rules shown in Table 1. Overall, the fact that Kim participated in a single challenge and used a single algorithm which performed joint nodule detection, segmentation and texture characterization may have also played a role. By having a single algorithm and objective, it becomes easier to tune parameters and obtain a final Fleischner probability class, whereas with independent parts, there are multiple probability outputs (nodule vs. non-nodule, segmentation, texture) which must be tuned and joined to obtain a final Fleischner class. In this way, it is possible that the method proposed by Kim would not have the best performance in terms of nodule detection when compared to Kaluva et al. but would still perform better in terms of Fleischner classification by consistently finding and performing robust characterization of the most important nodules for Fleischner classification.

In comparison to the interobserver variability, both Kaluva et al. and Kim showed statistically significantly lower performance at the train/validation stage (4 and 3 radiologists respectively). On the test stage however, the interobserver κ_w was lower, coming closer to the automatic classification, and only Kaluva et al. had statistically significantly different performance, and only to one out of five radiologists. This indicates that both approaches, and especially that by Kim, can capture the most important nodules, thus being able to predict patient follow-up with similar performance as radiologists. However, the limited data used for the McNemar's test on the test stage can have played a role in these results, limiting the applicability of this test.

6.2. *Challenge A - Nodule Detection*

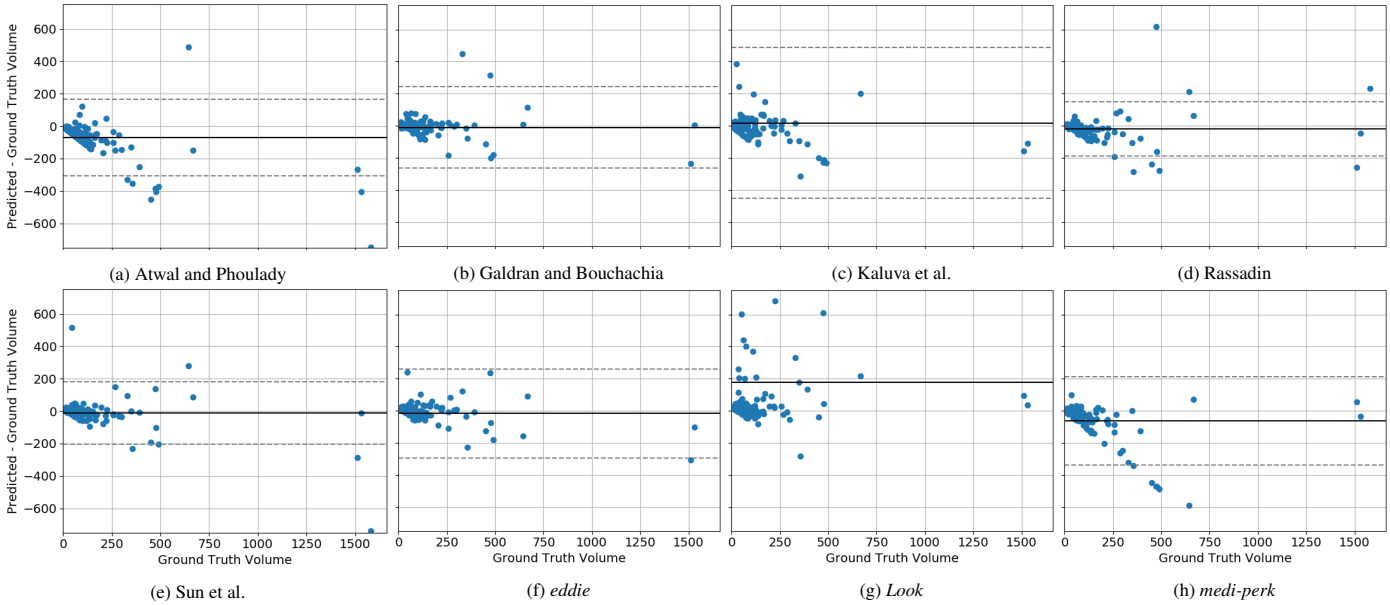
Focusing on the test stage results of challenge A, it can be seen that Kaluva et al.'s approach was more successful at retrieving true nodules at FP/scan levels below 1. While it is naturally difficult to say with certainty what factor caused the difference in performance between the two approaches, the difference is likely to be related to the detection module itself (rather than the FP reduction) and hyperparameter choices. While Kaluva et al. submitted around 19 nodule candidates per scan on both the train/validation and test stages, Katz et al. submitted below 8 nodule candidates per scan on the test stage compared to over 25 nodule candidates per scan on the train/validation stage. Such a low number of candidates per scan on the test stage is of course highly limiting as true nodules were likely excluded even before FP reduction. Another important factor to consider is the data used for training. While Katz et al. trained the detection module on LIDC-IDRI (LNDb was only used for training of the FP reduction module), Kaluva et al. used LIDC-IDRI, NLST and a private dataset as well, totalling over 2000 CTs, compared to 1018 by Katz et al.. In this way, the method trained by Kaluva et al. was trained with many more examples, leading



(a) Kaluva et al.

(b) Katz et al.

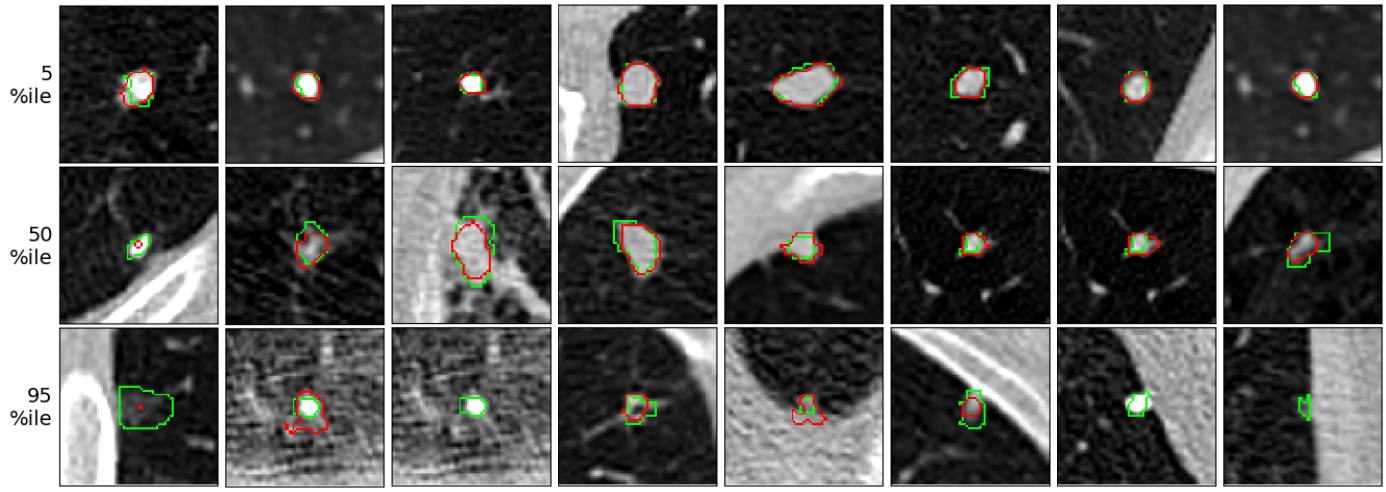
Fig. 3: Central axial view (51×51mm) of nodule candidate examples on test stage for both participants. Rows correspond to true positives (TP) and false positives (FP) and columns to each of the FP/scan levels considered for evaluation.



(a) Atwal and Phoulady (b) Galdran and Bouchachia (c) Kaluva et al. (d) Rassadin

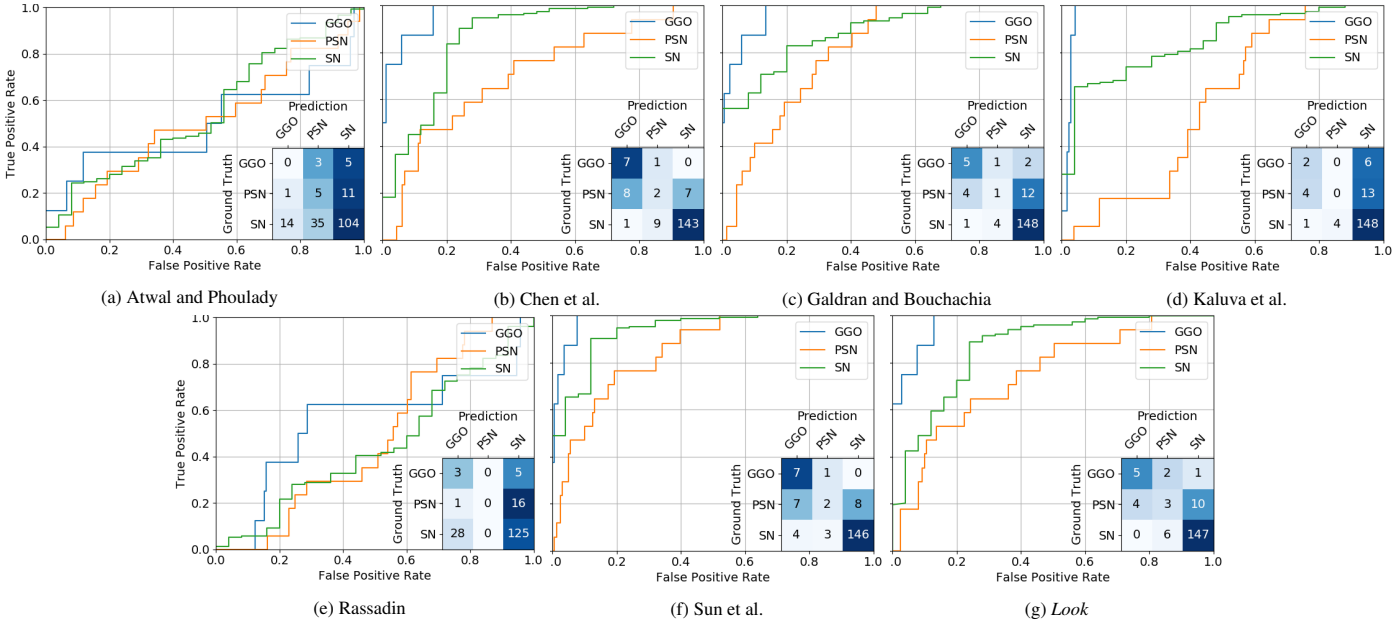
(e) Sun et al. (f) *eddie* (g) *Look* (h) *medi-perk*

Fig. 4: Nodule segmentation Bland-Altman plots obtained on test stage for each participant. Horizontal black line corresponds to mean volume difference and dashed gray lines to the 95% confidence interval. All volumes shown in mm^3. Absolute volume differences greater than $750mm^3$ not shown.



(a) Atwal and Phoulady (b) Galdran and Bouchachia (c) Kaluva et al. (d) Rassadin (e) Sun et al. (f) *eddie* (g) *Look* (h) *medi-perk*

Fig. 5: Central axial view (25×25mm) of nodule segmentation examples on test stage for all participants. Green corresponds to edges of ground truth segmentation and red to automatic segmentation. Rows correspond to examples at 5, 50 and 95 percentile values of s_B for each participant. Note that for Kaluva et al., *Look* and *medi-perk* no automatic segmentation is shown at 95 percentile because the predicted segmentation did not cross the central axial view shown.



(a) Atwal and Phoulady (b) Chen et al. (c) Galdran and Bouchachia (d) Kaluva et al.

(e) Rassadin (f) Sun et al. (g) *Look*

Fig. 6: Nodule texture characterization results obtained on test stage for each participant expressed through ROCs for each texture class and confusion matrix. ROCs for each texture class were obtained by considering each class as the positive class and the remaining classes as the negative.

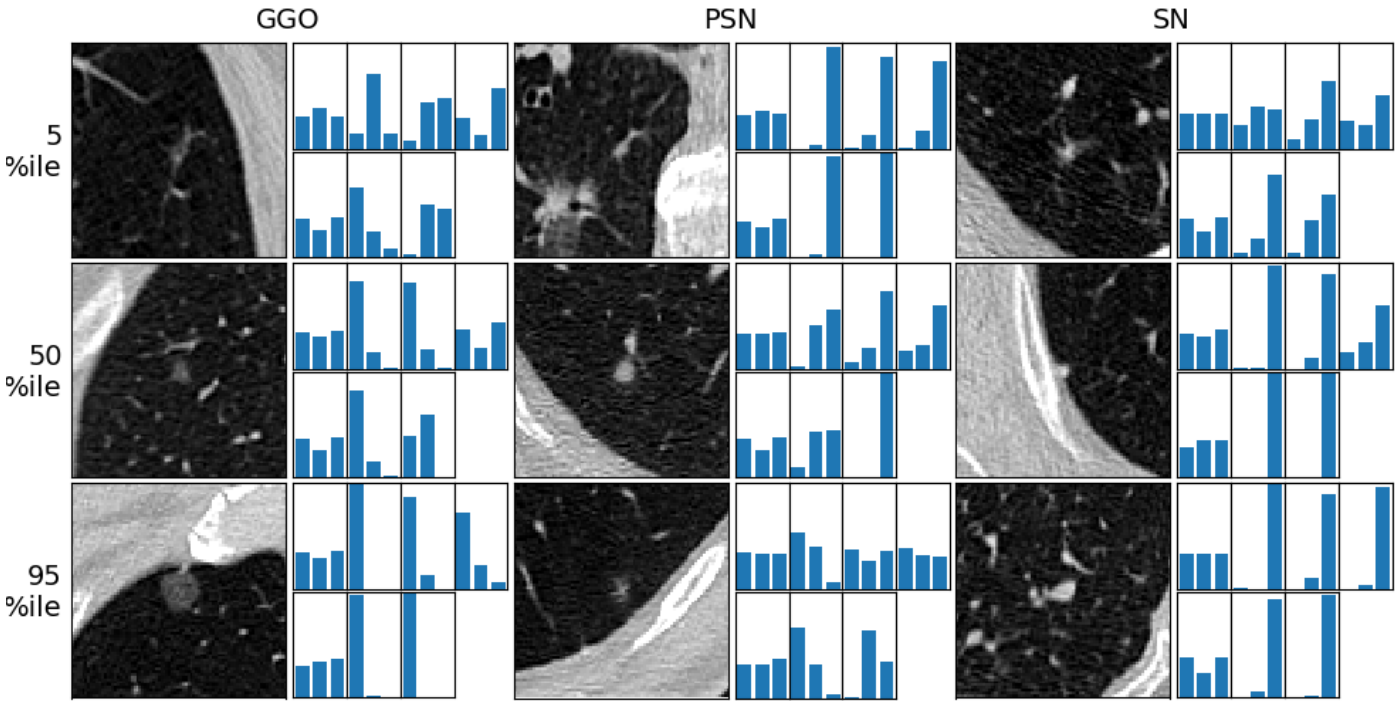


Fig. 7: Central axial view (51×51mm) of nodule texture characterization examples on test stage. Columns correspond to examples of each of the different nodule classes (GGO - Ground glass opacities, PSN - part solid nodules, SN - solid nodules) and rows correspond to examples at 5, 50 and 95 percentile values of the average ground truth class predicted probability averaged across all participants. Bar plots correspond to predicted probabilities for each of the three classes (GGO, PSN and SN from left to right) for each participant (Top row left to right: Atwal and Phoulady, Chen et al., Galdran and Bouchachia and Kaluva et al.; Bottom row left to right: Rassadin, Sun et al. and *Look*).

to a more efficient algorithm.

In terms of overall sensitivity, both approaches are well below previous nodule detection challenges such as LUNA16 where most approaches reach sensitivities above 0.9 at 1 FP/scan (Setio et al., 2017). This is in fact also the case for the approaches proposed by Kaluva et al. and Katz et al. who both report sensitivities of approximately 0.93 at 1 FP/scan on different subsets of the LIDC-IDRI dataset on which LUNA16 is based (Kaluva et al., 2020; Katz et al., 2020). As such, the lower performances obtained on LNDb are directly related to the CT data and annotations themselves. Firstly, there are significant differences in the CT data inclusion criteria, in specific on slice thickness. On LNDb, the maximum slice thickness is of 1.0mm, whereas on LIDC-IDRI the maximum slice thickness is of 5.0mm. Having a smaller slice thickness directly influences the manual annotation as there are more slices to inspect, but also allows for a more accurate representation of smaller nodules, being in this way conducive of a greater proportion of nodules <5mm being reported as shown in Pedrosa et al. (2019). Furthermore, these differences can have a detrimental impact on performance if a network is trained on LIDC-IDRI and is then deployed to LNDb without finetuning, as the network might not have adequate representations of the smaller slice thicknesses observed on LNDb. Secondly, the fact that annotations on LNDb were made in a single blind fashion without revision of findings by other radiologists (in contrast to LIDC-IDRI where revision was performed) has an influence in the level of agreement among radiologists. Whereas on LIDC-IDRI the level of agreement can be solely attributed to decision error, i.e. deciding if each finding is a nodule or not, on LNDb, the level of agreement compounds the decision and fixation errors, i.e. the process of actually finding a nodule in the 3D CT image. Furthermore, the higher proportion of nodules <5mm can increase decision error, as the size of the finding can make decision more difficult (Gierada et al., 2008). The fact that in LNDb the number of radiologists that annotated each image is variable also changes the meaning of the agreement level in a given nodule, as an obvious nodule will have an agreement level of 1 if the corresponding CT was annotated by a single radiologist. These inherent differences in annotation and agreement

mean that the criteria used in LUNA16 for computing performance (limiting to nodules with agreement level of 3 or higher) could not be used, which makes a direct comparison between LUNA16 and LNDb results not feasible. However, the single blind annotations performed for LNDb and resulting interobserver variability are more representative of clinical reality.

As shown in Figure 2, interobserver variability on LNDb was around 1 FP/scan, with sensitivities close to 0.5 for agreement level 1 and 0.65 and 0.9 for agreement level 2 for the train/validation and test stages respectively. Both Kaluva et al. and Katz et al. scored below the interobserver sensitivity at 1 FP/scan, achieving the same sensitivity on the test stage at approximately 2 FP/scan for agreement level 1 and above 4 FP/scan for agreement level 2. As such, it can be hypotethized that further improvements can be possible with further refinement of the detection algorithms. In fact, Kaluva et al. did not use the LNDb data for training and Katz et al. used it for training the FP reduction module only. As such, network finetuning using LNDb data could yield significant improvement bringing performance closer to interobserver levels.

6.3. Challenge B - Nodule Segmentation

While Sun et al. and Rassadin obtained the best results on the train/validation and test stages on challenge B, neither of these participants obtained the best results in terms of segmentation performance measures, which were obtained by *Look* and *eddie* on the train/validation and test stages respectively. *Look*'s results on the train/validation stage, which were statistically significantly different from all other submissions, were however overfit to the train/validation data, and test stage performance was poor. *eddie*'s results on the test stage, which were statistically significantly different from Rassadin and Sun et al. in terms of *MAD*, failed to achieve the best s_B as they have a higher r^* and slightly higher bias and standard deviation. This is likely a consequence of worst segmentation predictions for larger nodules as shown in Figure 4, which will significantly degrade volume performance measures. While it can seem counterintuitive that the best results in terms of segmentation performance measures did not correspond to the winners of challenge B, the use of volume performance measures can give additional insight into segmentation performance and is particularly meaningful if, as in this case, nodule volume is the clinical endpoint in sight and not the actual segmentation of the nodule borders.

In terms of methodology, all reported methods were based on 3D U-Net architecture though there are significant differences in terms of the training details. Both Rassadin and Sun et al. performed joint training of segmentation and texture, having achieved the best results in terms of s_B. Joint segmentation and texture classification might thus have been a deciding factor by avoiding overfitting to the training data and generating more representative features. In terms of input size and voxelsize, all participants except Kaluva et al. used the submission input size and voxelsize of 80×80×80 and 0.6375mm. This submission input size and voxelsize was chosen as it was large enough for the representation of the segmentation of all nodules and used a voxelsize close to the average resolution of the CT axial slices, thus avoiding downsampling of the nodules' segmentation. It can be seen that the choice of Kaluva et al. to increase input size and voxelsize, which could have been beneficial by giving more context of the nodule surroundings, led to below average segmentation performance. However, the fact that a larger voxelsize (1mm) was used by Kaluva et al., may have been responsible for these worst results as well and the influence of each parameter separately would have to be investigated. The most different parameter among all methods is the HU window used to truncate the input. Changing the HU window is a tool commonly used by radiologists to better visualize nodules and nodule boundaries, which explains the focus of participants on this factor. However, no particular trend can be observed between the performance and HU window used, and it is expected that a 3D U-Net will be able to learn the most efficient representation independently of the input HU window (within reasonable limits) and this is confirmed by the fact that the winning method by Rassadin did not perform any HU window truncation. Finally, only two participants, Kaluva et al. and Sun et al., used data external to LNDb for training and only Sun et al. performed pretraining. However, comparing Rassadin and Sun et al., which

used similar approaches and obtained similar results, the added training data and pretraining methodology does not seem to be particularly impactful to the final performance.

One of the notable aspects of challenge B is the fact that several methods outperform the interobserver variability. On the test stage, Galdran and Bouchachia, Rassadin and *eddie* have statistically significantly lower J^*, MAD and/or HD and report a lower absolute volume bias. This is indicative of the advanced state of the art in segmentation, in spite of the complexity of the task. While further advances in overall performance are always possible, their clinical meaningfulness becomes lower, and a more critical performance evaluation must be performed in the future. It has been shown that GGO and PSN are more difficult to segment (Aresta et al., 2019), both due to their lower constrast to the lung parenchyma and due to their lower frequency during training, and an ideal segmentation algorithm must be able to cope with these challenges. Furthermore, the location of the nodule is also known to be a critical factor for performance and this is clearly shown in Figure 5 where 5 of the 95%ile examples are juxta-pleural, in contrast to only 1 of the 5%ile examples. Greater context of the surrounding image beyond the nodule boundaries can play a key role in improving segmentation for these types of nodules.

6.4. Challenge C - Nodule Texture Characterization

On Challenge C, Chen et al. obtained the best test stage κ_w, whereas Atwal and Phoulady obtained the best train/validation stage κ_w. However, Atwal and Phoulady's results on the train/validation stage, which were statistically significantly different from all other submissions, were overfit to the train/validation data, and test stage performance was poor. On the test stage, the only submissions statistically significantly different from Chen et al. were, those by Atwal and Phoulady and Rassadin, though this is likely due to the small amount of data on which the McNemar's test relies upon. This is also the case for the lack of statistically significant differences with the interobserver variability, and it can be seen that on both train/validation and test stages, no submission other than that by Atwal and Phoulady reaches the interobserver variability.

As expected, most methods exhibit high performance in the classification of solid nodules as shown in Figure 6 which is a consequence of the overwhelming majority of this class. The class showing the lowest performance is the part solid nodules, not only due to the low number of train/test examples but also likely due to the fact that, because it is the middle class, it exhibits a higher variability of appearances as shown in Figure 7. The top example of the PSN column illustrates this issue as it shows a nodule annotated as PSN but that was classified as solid by six of the participants and which is not significantly different in appearance from the top example on the SN column. The extent of misclassification of the PSN class is clearly shown on the confusion matrixes of Figure 6, where the participant with highest accuracy for this class (Atwal and Phoulady) only classified correctly 5 of the 17 nodules of this class.

Regarding the proposed methods, there is a much higher variability in comparison to Challenge B in terms of architectures, though all techniques rely on deep learning. In terms of HU window and voxelsize, the same trends of Challenge B were observed. Both Rassadin and Sun et al. used similar approaches, performing joint segmentation and texture classification using the features extracted by the segmentation encoder, with the difference that Sun et al. also used the gray value distribution as features for classification. Nevertheless, the results between these two participants are radically different and the poor performance obtained by Rassadin on both train/validation and test suggests that there were issues with model convergence or generating valid predictions. In terms of input size, Chen et al. and Galdran and Bouchachia were the only participants using 2D slices rather than a 3D volume and both had good performance on the train/validation and test stages. While it might be tempting to assume that the use of 3D volumes will give greater information and thus greater performance, two important factors must be taken into account. Firstly, radiologists typically use the axial slice for CT visualization, using the sagittal and coronal slices only in rare occasions and mostly during nodule detection (i.e. to determine whether a certain structure is in fact a nodule). In that sense, 3D features are not mostly used

by radiologists for texture characterization. Secondly, using a 2D approach has the additional advantages that 2D networks have, in general, less features to train and that different slices of the same nodule can be used for training as an augmentation strategy. This can be particularly important when dealing with smaller datasets such as LNDb. The use of additional datasets can also be important, and this strategy was applied by Chen et al. but also by Kaluva et al. and Sun et al.. Finally, Chen et al. started from a pretrained model from ImageNet, while Sun et al. used Models Genesis as a pretraining strategy.

Comparing the top four participants (Chen et al., *Look*, Sun et al. and Galdran and Bouchachia) and corresponding ROCs on Figure 6, it can be seen that while the final score according to κ_w were similar, there are significant differences in performance. Sun et al. in particular, has a much higher area under the curve for part-solid nodules than any other participant. However, these differences become diluted in the κ_w measure used, which only takes into account the highest probability for any given nodule. The low number of GGOs and PSNs makes it so that κ_w is extremely sensitive to misclassifications outside the neighboring class (solid nodules predicted as GGOs or viceversa) as highlighted by comparing the confusion matrixes of Chen et al. and Sun et al.. As such, the use of additional evaluation measures, such as area under the curve for each class, could be important in the future to more robustly evaluate classification performance, especially for problems with little and/or imbalanced data.

6.5. Limitations

While promising results have been shown in this study, there are limitations to this study which must be taken into account for an adequate analysis of the results.

Regarding the LNDb dataset, while 294 CTs is a moderate quantity which is associated to a significant amount of annotation and curation work, it remains a relatively small dataset, especially after partitioning into train/validation and test sets, and taking into account the imbalance of different nodule and Fleischner classes. While most, if not all, participants increased the training data by adding public datasets, this limitation may have played a role in the successful training and generalization capabilities of the methods proposed, which mostly rely on deep learning methodologies. Furthermore, and as seen throughout this Section, the relative small size of the test set has limited the interpretation of the results, especially for the Fleischner and texture classification where class imbalance was a serious concern. As such, further increasing the available data and its diversity would enable finer training and methoodologies as well as more detailed evaluations and conclusion, ultimately leading to better performing algorithms. Regarding the annotations of the LNDb dataset, while having access to a single blind reading of the CTs gives a better understanding of interobserver variability, it would be extremely useful to have a second reading where radiologists exchange annotations. In that way, a more accurate ground truth would be obtained, to which the typical single blind reading of a radiologist could be compared. However, running such a reading for even the modest size of 294 CTs and five radiologists was simply not achievable in a reasonable amount of time/effort.

Regarding the challenge itself, the greatest limitation is the small number of submissions to the main challenge and challenge A. While the limited submissions on challenge A may have been influenced by the overlap with previous challenges on CT nodule detection, such as LUNA16 (Setio et al., 2017) and ANODE (Van Ginneken et al., 2010), the complexity of the the main challenge and challenge A may have also played a role. Challenges B and C were the most straightforward and received the highest number of submissions. The main challenge was particularly complex as it implied solving challenges A, B and C to successfully apply the Fleischner rules or performing a direct classification at CT level which presents significant challenges due to the size of the data. Nevertheless and in spite of these limitations, the fact that the LNDb challenge will remain online and available for submissions in the foreseeable future will ensure that the significance of the LNDb Challenge and dataset is liable to increase with time.

7. Conclusion

The LNDb challenge was organized to promote research and the benchmarking of automatic algorithms on automatic lung cancer patient management, overcoming previous challenges which were focused on a specific aspect of lung cancer CT screening/management. In this way, the LNDb Challenge encompassed nodule detection, segmentation and texture characterization with the final goal of automatic patient management according to the 2017 Fleischner guidelines. While significant research efforts have been developed in this area so far, most tasks still require improvement in order to achieve radiologist-level performance as shown in this study. Nodule detection is particularly challenging due to the size of the data and the many structures that resemble nodules in CT scans and the class imbalance and subjective nature of nodule texture are the main challenges identified for robust nodule characterization and Fleischner classification. Nevertheless, the advance of image analysis methodologies and future submissions to LNDb will certainly pave the way for more efficient, better performing methodologies, which will one day be able to become a valuable second opinion to radiologists in the screening and management of lung cancer.

Acknowledgments

This work was financed by the European Regional Development Fund (ERDF) through the Operational Programme for Competitiveness - COMPETE 2020 Programme and by National Funds through the Portuguese Funding agency, FCT - Fundação para a Ciência e Tecnologia within projects PTDC/EEI-SII/6599/2014 (POCI-01-0145-FEDER-016673) and UIDB/50014/2020. Guilherme Aresta is funded by the FCT grant contract SFRH/BD/120435/2016. Carlos Ferreira is funded by the FCT grant contract SFRH/BD/146437/2019.

References

Aresta, G., Araújo, T., Jacobs, C., van Ginneken, B., Cunha, A., Ramos, I., Campilho, A., 2018. Towards an automatic lung cancer screening system in low dose computed tomography, in: Image Analysis for Moving Organ, Breast, and Thoracic Images. Springer, pp. 310–318.

Aresta, G., Jacobs, C., Araújo, T., Cunha, A., Ramos, I., van Ginneken, B., Campilho, A., 2019. iW-Net: an automatic and minimalistic interactive lung nodule segmentation deep network. Scientific reports 9, 1–9.

Armato, S.G., McLennan, G., Bidaut, L., McNitt-Gray, M.F., Meyer, C.R., Reeves, A.P., Zhao, B., Aberle, D.R., Henschke, C.I., Hoffman, E.A., et al., 2011. The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans. Medical physics 38, 915–931.

Atwal, G., Phoulady, H.A., 2020. Automatic lung cancer follow-up recommendation with 3D deep learning, in: International Conference Image Analysis and Recognition, Springer. pp. 0–0.

Causey, J.L., Zhang, J., Ma, S., Jiang, B., Qualls, J.A., Politte, D.G., Prior, F., Zhang, S., Huang, X., 2018. Highly accurate model for prediction of lung nodule malignancy with ct scans. Scientific reports 8, 1–12.

Chen, Y., Li, J., Xiao, H., Jin, X., Yan, S., Feng, J., 2017. Dual path networks, in: Advances in neural information processing systems, pp. 4467–4475.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: learning dense volumetric segmentation from sparse annotation, in: International conference on medical image computing and computer-assisted intervention, Springer. pp. 424–432.

Clevert, D.A., Unterthiner, T., Hochreiter, S., 2015. Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289 .

Dey, R., Lu, Z., Hong, Y., 2018. Diagnostic classification of lung nodules using 3d neural networks, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE. pp. 774–778.

Ding, J., Li, A., Hu, Z., Wang, L., 2017. Accurate pulmonary nodule detection in computed tomography images using deep convolutional neural networks, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 559–567.

Edwards, A.L., 1948. Note on the "correction for continuity" in testing the significance of the difference between correlated proportions. Psychometrika 13, 185–187.

Ferreira, C.A., Cunha, A., Mendonça, A.M., Campilho, A., 2018. Convolutional neural network architectures for texture classification of pulmonary nodules, in: Iberoamerican Congress on Pattern Recognition, Springer. pp. 783–791.

Galdran, A., Bouchachia, H., 2020. Residual networks for pulmonary nodule segmentation and texture characterization, in: International Conference Image Analysis and Recognition, Springer. pp. 0–0.

Galdran, A., Chelbi, J., Kobi, R., Dolz, J., Lombaert, H., ben Ayed, I., Chakor, H., 2020. Non-uniform label smoothing for diabetic retinopathy grading from retinal fundus images with deep neural networks. Translational Vision Science & Technology 9, 34–34.

Gierada, D.S., Pilgram, T.K., Ford, M., Fagerstrom, R.M., Church, T.R., Nath, H., Garg, K., Strollo, D.C., 2008. Lung cancer: interobserver agreement on interpretation of pulmonary findings at low-dose CT screening. Radiology 246, 265–272.

Glorot, X., Bordes, A., Bengio, Y., 2011. Deep sparse rectifier neural networks, in: 14th international conference on artificial intelligence and statistics, pp. 315–323.

Golan, R., Jacob, C., Denzinger, J., 2016. Lung nodule detection in ct images using deep convolutional neural networks, in: 2016 International Joint Conference on Neural Networks (IJCNN), IEEE. pp. 243–250.

Gonçalves, L., Novo, J., Cunha, A., Campilho, A., 2018. Learning lung nodule malignancy likelihood from radiologist annotations or diagnosis data. Journal of Medical and Biological Engineering 38, 424–442.

Han, H., Li, L., Han, F., Song, B., Moore, W., Liang, Z., 2014. Fast and adaptive detection of pulmonary nodules in thoracic ct images using a hierarchical vector quantization scheme. IEEE journal of biomedical and health informatics 19, 648–659.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 .

Isensee, F., Petersen, J., Klein, A., Zimmerer, D., Jaeger, P.F., Kohl, S., Wasserthal, J., Koehler, G., Norajitra, T., Wirkert, S., et al., 2018. nnu-net: Self-adapting framework for u-net-based medical image segmentation. arXiv preprint arXiv:1809.10486 .

Kaluva, K.C., Vaidhya, K., Chunduru, A., Tarai, S., Nadimpalli, S.P.P., Vaidya,

S., 2020. An automated workflow for lung nodule follow-up recommendation using deep learning, in: International Conference Image Analysis and Recognition, Springer. pp. 0–0.

Katz, O., Presil, D., Cohen, L., Schwartzbard, Y., Kashani, S., Hoch, S., 2020. Pulmonary-nodule detection using an ensemble of 3d SE-ResNet18 and DPN68 models, in: International Conference Image Analysis and Recognition, Springer. pp. 0–0.

MacMahon, H., Naidich, D.P., Goo, J.M., Lee, K.S., Leung, A.N., Mayo, J.R., Mehta, A.C., Ohno, Y., Powell, C.A., Prokop, M., et al., 2017. Guidelines for management of incidental pulmonary nodules detected on CT images: from the Fleischner Society 2017. Radiology 284, 228–243.

McNitt-Gray, M.F., Armato III, S.G., Meyer, C.R., Reeves, A.P., McLennan, G., Pais, R.C., Freymann, J., Brown, M.S., Engelmann, R.M., Bland, P.H., et al., 2007. The lung image database consortium (LIDC) data collection process for nodule detection and annotation. Academic radiology 14, 1464–1474.

Messay, T., Hardie, R.C., Rogers, S.K., 2010. A new computationally efficient cad system for pulmonary nodule detection in ct imagery. Medical image analysis 14, 390–406.

Murphy, K., van Ginneken, B., Schilham, A.M., De Hoop, B., Gietema, H., Prokop, M., 2009. 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, 757–770.

Pedrosa, J., Aresta, G., Ferreira, C., Rodrigues, M., Leitão, P., Carvalho, A.S., Rebelo, J., Negrão, E., Ramos, I., Cunha, A., et al., 2019. Lndb: A lung nodule database on computed tomography. arXiv preprint arXiv:1911.08434 .

Rassadin, A., 2020. Deep residual 3D U-Net for joint segmentation and texture classification of nodules in lung, in: International Conference Image Analysis and Recognition, Springer. pp. 0–0.

Reeves, A.P., Kostis, W.J., 2000. Computer-aided diagnosis of small pulmonary nodules, in: Seminars in Ultrasound, CT and MRI, Elsevier. pp. 116–128.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in neural information processing systems, pp. 91–99.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer. pp. 234–241.

Setio, A.A.A., Traverso, A., De Bel, T., Berens, M.S., van den Bogaard, C., Cerello, P., Chen, H., Dou, Q., Fantacci, M.E., Geurts, B., et al., 2017. 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.

Shen, S., Han, S.X., Aberle, D.R., Bui, A.A., Hsu, W., 2019. An interpretable deep hierarchical semantic convolutional neural network for lung nodule malignancy classification. Expert Systems with Applications 128, 84–95.

Siegel, R.L., Miller, K.D., Jemal, A., 2019. Cancer statistics, 2019. CA: a cancer journal for clinicians .

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 .

Spitzer, R.L., Cohen, J., Fleiss, J.L., Endicott, J., 1967. Quantification of agreement in psychiatric diagnosis: A new approach. Archives of General Psychiatry 17, 83–87.

Sun, Z., Jia, Y., Men, X., Tian, Z., 2020. 3DCNN for pulmonary nodule segmentation and classification, in: International Conference Image Analysis and Recognition, Springer. pp. 0–0.

The National Lung Screening Trial Research Team, 2011. Reduced lung-cancer mortality with low-dose computed tomographic screening. New England Journal of Medicine 365, 395–409.

Van Ginneken, B., Armato III, S.G., de Hoop, B., van Amelsvoort-van de Vorst, S., Duindam, T., Niemeijer, M., Murphy, K., Schilham, A., Retico, A., Fantacci, M.E., et al., 2010. Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the anode09 study. Medical image analysis 14, 707–722.

Van Rikxoort, E., Van Ginneken, B., 2011. Automatic segmentation of the lungs and lobes from thoracic ct scans, in: Proc. 4th Int. Workshop Pulmonary Image Anal, pp. 261–268.

Wang, G., Li, W., Aertsen, M., Deprest, J., Ourselin, S., Vercauteren, T., 2019. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. Neurocomputing 338, 34–45.

Wang, S., Zhou, M., Liu, Z., Liu, Z., Gu, D., Zang, Y., Dong, D., Gevaert, O., Tian, J., 2017. Central focused convolutional neural networks: Developing a data-driven model for lung nodule segmentation. Medical image analysis 40, 172–183.

Wu, B., Zhou, Z., Wang, J., Wang, Y., 2018. Joint learning for pulmonary nodule segmentation, attributes and malignancy prediction, in: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), IEEE. pp. 1109–1113.

Wu, Y., He, K., 2018. Group normalization, in: Proceedings of the European conference on computer vision (ECCV), pp. 3–19.

Zagoruyko, S., Komodakis, N., 2016. Wide residual networks. arXiv preprint arXiv:1605.07146 .

Zhou, Z., Sodha, V., Siddiquee, M.M.R., Feng, R., Tajbakhsh, N., Gotway, M.B., Liang, J., 2019. Models genesis: Generic autodidactic models for 3d medical image analysis, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer. pp. 384–393.