Computer Methods and Programs in Biomedicine 206 (2021) 106111

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

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

A simplified cluster model and a tool adapted for collaborative labeling of lung cancer CT scans

S.P. Morozov[a], V.A. Gombolevskiy[a], A.B. Elizarov[a], M.A. Gusev[a,d], V.P. Novik[a], S.B. Prokudaylo[a], A.S. Bardin[a], E.V. Popov[a], N.V. Ledikhova[a], V.Y. Chernina[a], I.A. Blokhin[a], A.E. Nikolaev[a], R.V. Reshetnikov[a,b], A.V. Vladzymyrskyy[a], N.S. Kulberg[a,c,*]

[a] *Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department, Petrovka str., 24, Moscow, 127051, Russia*
[b] *Institute of Molecular Medicine, Sechenov First Moscow State Medical University, Trubetskaya str. 8-2, Moscow, 119991, Russia*
[c] *Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Vavilova str., 44/2, Moscow, 119333, Russia*
[d] *Federal State Budgetary Educational Institution of Higher Education "Moscow Polytechnic University", Tverskaya str., 11, Moscow, 125993, Russia*

ARTICLE INFO

Article history:
Received 1 September 2020
Accepted 7 April 2021

Keywords:
Lung nodule
Computed tomography (CT)
Computer-aided diagnosis (CAD)
Machine learning (ML)
Training dataset
Crowdsourcing
Weak labeling

ABSTRACT

Background and objective: Lung cancer is the most common type of cancer with a high mortality rate. Early detection using medical imaging is critically important for the long-term survival of the patients. Computer-aided diagnosis (CAD) tools can potentially reduce the number of incorrect interpretations of medical image data by radiologists. Datasets with adequate sample size, annotation, and truth are the dominant factors in developing and training effective CAD algorithms. The objective of this study was to produce a practical approach and a tool for the creation of medical image datasets.

Methods: The proposed model uses the modified maximum transverse diameter approach to mark a putative lung nodule. The modification involves the possibility to use a set of overlapping spheres of appropriate size to approximate the shape of the nodule. The algorithm embedded in the model also groups the marks made by different readers for the same lesion. We used the data of 536 randomly selected patients of Moscow outpatient clinics to create a dataset of standard-dose chest computed tomography (CT) scans utilizing the double-reading approach with arbitration. Six volunteer radiologists independently produced a report for each scan using the proposed model with the main focus on the detection of lesions with sizes ranging from 3 to 30 mm. After this, an arbitrator reviewed their marks and annotations.

Results: The maximum transverse diameter approach outperformed the alternative methods (3D box, ellipsoid, and complete outline construction) in a study of 10,000 computer-generated tumor models of different shapes in terms of accuracy and speed of nodule shape approximation. The markup and annotation of the CTLungCa-500 dataset revealed 72 studies containing no lung nodules. The remaining 464 CT scans contained 3151 lesions marked by at least one radiologist: 56%, 14%, and 29% of the lesions were malignant, benign, and non-nodular, respectively. 2887 lesions have the target size of 3–30 mm. Only 70 nodules were uniformly identified by all the six readers. An increase in the number of independent readers providing CT scans interpretations led to an accuracy increase associated with a decrease in agreement. The dataset markup process took three working weeks.

Conclusions: The developed cluster model simplifies the collaborative and crowdsourced creation of image repositories and makes it time-efficient. Our proof-of-concept dataset provides a valuable source of annotated medical imaging data for training CAD algorithms aimed at early detection of lung nodules. The tool and the dataset are publicly available at https://github.com/Center-of-Diagnostics-and-Telemedicine/FAnTom.git and https://mosmed.ai/en/datasets/ct_lungcancer_500/, respectively.

1. Introduction

Lung cancer, a highly invasive and rapidly metastasizing disease, is the most common type of cancer associated with a poor progno-

[*] Corresponding author at: Research and Practical Clinical Center for Diagnostics and Telemedicine Technologies of the Moscow Health Care Department, Petrovka str., 24, Moscow 127051, Russia.
E-mail address: kulberg@npcmr.ru (N.S. Kulberg).

https://doi.org/10.1016/j.cmpb.2021.106111
0169-2607/© 2021 Elsevier B.V. All rights reserved.

sis [1]. According to Bray et al., in 2018 [2], it was the leading cause of cancer deaths worldwide (18.4%), followed by stomach (8.2%), liver (8.2%), and breast (6.6%) cancer . Although a number of new targeted agents and immunotherapies are being developed, early detection and treatment are still the best options for the long-term survival of lung cancer patients [3,4]. This requires routine monitoring of high-risk subjects using computed tomography (CT) and involves examination of a huge number of CT scans by radiologists [5]. Computer-aided diagnosis (CAD) tools based on machine learning (ML) models are intended to assist the radiologists by marking suspicious features on chest images aiding the human inspection.

The cornerstone of developing and improving accurate and computationally efficient ML models is the availability of high-quality training and testing datasets. The main datasets used in lung cancer research are the combined database of the Lung Image Database Consortium and the Image Database Resource Initiative (LIDC/IRDI) [6], the LUNA16 subset of LIDC/IDRI database [7], the dataset provided by the LUNGx challenge organized by SPIE, the American Association of Physicists in Medicine (AAPM), and the National Cancer Institute (NCI) [8], the Lung Test Images from Motol Environment (Lung TIME) database [9], ANODE09 database [10], the database of Lung CT Imaging Signs (LISS) [11], and the data from National Lung Screening Trial (NLST) of NCI [12].

Currently, most large datasets for lung cancer research are created from images acquired in screening trials and therefore consist of low-dose CT (LDCT) scans. Unsurprisingly, the most notable achievements in performance of ML models are made in this area [13,14]. However, LDCT has its limitations [15], and for some scenarios, the use of standard-dose CT is preferable. Several studies report that standard-dose CT images provide data for radiomics analysis that can be used for early detection of metastases [16–18]. Unfortunately, these studies rely on non-public datasets of limited size, which does not allow the fine-tuning of the proposed methods. The insufficient availability of large amounts of accurately annotated training data currently is a bottleneck of this line of research.

There is a variety of software tools developed for medical image annotation [19–23]. They enable partial or full automation of the labeling process, but the interpretation of radiological data still depends on human intelligence. Crowdsourcing platforms have performed well in cost-effective large-scale image annotation [24]; however, they have limitations as the correct reading of CT scans requires special training and experience [25]. Weak labeling approaches (for example, free-text radiology reports [26], bounding boxes, or outlier correction with the use of a weakly labeled atlas [27]) are proposed to reduce the workload of medical experts.

We propose an open-source tool adapted for collaborative multitenant annotation of CT scan datasets, available at https://github.com/Center-of-Diagnostics-and-Telemedicine/FAnTom.git. The tool is based on a cluster model for nodule localization. The model's main features are the tolerance to slight differences in interpretations of individual readers and the ability to describe complex-shaped lesions with low effort. Using the double-reading approach with arbitration for ground truth annotation, we have created CTLungCa-500, a publicly available "proof-of-concept" dataset of thoracic standard-dose CT scans, consisting of 536 cases of patients with a high risk of lung cancer.

2. Materials and methods

2.1. Patient data

The Mandatory Health Insurance System of the Russian Federation provides free health services for everyone who resides permanently or temporarily in Russia. Federal Law No. 326-FZ regulates the collection of personal data relating to all diagnoses, outcomes, forms, duration, and scope of medical care. Clinical data are stored in the Unified Medical Information Analysis System (UMIAS), and the corresponding medical images are stored in the Unified Radiology Information System (URIS). For some procedures, including CT, patients sign informed consent to use their anonymized data for scientific purposes. For this multi-center study, we retrospectively collected the URIS/UMIAS data of patients of Moscow (Russia) outpatient clinics, aged 50 to 75 years, who underwent diagnostic standard-dose chest CT imaging between January 2015 and December 2017, according to an attending physician's referral due to suspected lung cancer. The initial dataset contained 3897 thoracic CT scans. Of these, we randomly selected a subset of 550 scans from different patients for mock-up annotation and markup using the `random` module of Python 3.8.2 [28]. Our goal was to create a dataset of minimally sufficient size for the training of a classifier ML model. The size of the dataset was based on the results of Figueroa et al., according to which it takes between 80 to 560 annotated samples to achieve the desired performance of a classifier algorithm [29]. Fourteen CT scans were excluded due to non-compliance with the patient age criteria or imaging protocol requirements as some studies were performed using a low-dose or pediatric CT protocol. While creating the database, all protected health information of the patients was removed from the DICOM headers using in-house medical research anonymization software.

CT scans were acquired according to CT scanners manufacturers' protocols. The recommended scanning parameters for standard-size patients (height, 170 cm; weight, 70 kg) were as follows: automatic tube current modulation at the mean potential of 120 kV, 350-mm field of view, slice thickness $\leq$ 1.5 mm [30], spacing between adjacent slices $\leq$ slice thickness. Scans were acquired with subjects in the supine position, from the diaphragm to the apex of the lung within a single breath-hold. Reconstruction kernels were manufacturer-specific. Toshiba scanners: FC50, FC51, FC52, FC53, FC07 for lung tissue, and FC07, FC08, FC09, FC17, FC18 for soft tissue. Siemens: B70, B75, and B80. Philips: "Y-Sharp" and "LUNG" for lung tissue, "SOFT" for soft tissue. General Electrics scanner models: "LUNG" for lung tissue, "SOFT" for soft tissue.

2.2. Annotation and markup

Interpretation of CT scans by radiologists is subjective and error-prone [31,32], which can be compensated by the double-reading approach [33]. When the costs of false-positive and false-negative errors are equally high, an arbitration of initial readings proved to provide the optimal accuracy [34]. Note that arbitration is effective only if the initial readers make different mistakes [34]. Therefore, the number of readers can directly affect the accuracy of markup and annotation. According to Herman and Hessel, a given false-positive finding is unlikely to be discovered by more than one radiologist. However, a large proportion of false-negative errors are made by two and more readers [35]. Our study involved two groups of radiologists. The first group, consisting of 15 volunteer specialists with an experience of 2 to 10+ years, performed the initial reading. To reduce the probability of omissions on CT scans, six randomly chosen radiologists from this group independently read every included case. They were instructed to limit the markup to five lesions with sizes ranging from 3 to 30 mm per CT scan and ignore calcified and perifissural lung nodules. This cutoff was based on the results of the NELSON trial, which showed that the risk of primary cancer increases with the count of nodules $\leq$ 4, but decreases for patients with five or more nodules [36]. A single representative of the second group, which included three radiologists with an experience of 10+ years, provided the arbitration of the six reader reports.

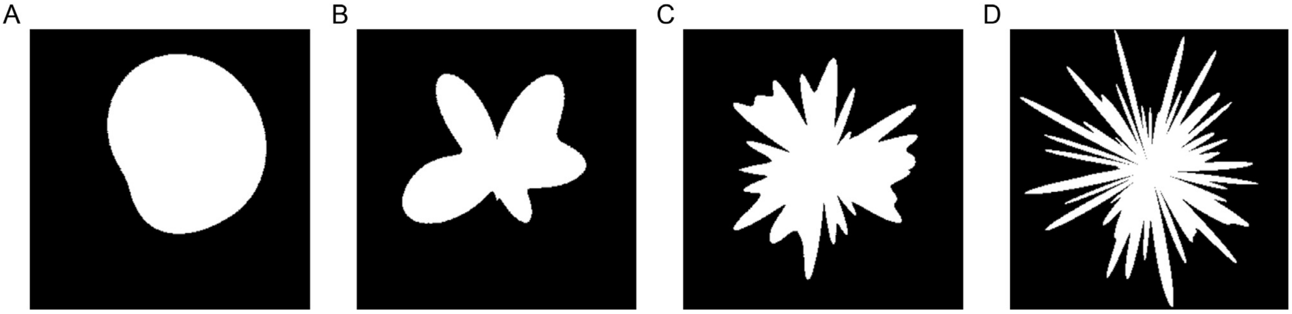

Fig. 1. Examples of tumor models. *A-D*: cross-sections of tumors with 20, 80, 1960 and 7850 growth lines, respectively.

2.2.1. Evaluation of accuracy and efficiency of different markup strategies

Several strategies exist to mark lung nodules. All of them require indication of the center-of-mass location for an abnormality but differ in the approach to describe its shape and size. The nodule geometry can be approximated by a 3D box, sphere, ellipsoid, or a complete outline. We compared the accuracy and efficiency of the approaches in mock-up measurements of computer-generated 3D lung nodule models to choose the optimal markup strategy. The models were built on the assumption that tumors evolve from a single initial cell [37], and the growth rate in different directions is heterogeneous, resulting in a spiculated appearance. For each tumor model, the number of independent growth lines was randomly assigned from 1 (spherical mass) to 7850 (highly spiculated mass). Each line had its own growth rate ranging from 0.1 to 1.0 (Fig. 1).

The quality of tumor shape approximation using different approaches was evaluated for 10,000 models of varying complexity. For this, we fit the models into 3D boxes, spheres, or ellipsoids of the minimum size, which allowed to include all points of the object using in-house FAnTom (Find Anomalies in Tomography) software described in Kulberg et al. [38]. The approximation quality was evaluated with the Sørensen–Dice coefficient:

$$d = 2 * \frac{|X \cap Y|}{|X| + |Y|}, \tag{1}$$

where X and Y are two sets, $|X|$ is the number of elements in the set X, $|X \cap Y|$ is the number of elements that are common to both sets.

Ten tumor models were used for a comparative efficiency study of different markup strategies. Each model was approximated with a 3D box, sphere, ellipsoid, or an outline was constructed around the tumor using either ITK-Snap [20], 3D Slicer [21], or the FAnTom software. The time spent on each shape approximation was recorded and used to calculate the mean value for each measurement strategy.

2.2.2. Architecture of the FAnTom software

The FAnTom software consists of three modules: web server, web service, and client application. The web service controls the markup process workflow. Web service instances are Linux applications that run in Docker containers managed by the web server. The web server is responsible for interactions with the client application, user authentification, and managing database and PACS connections. The web server runs in the JVM (Java Virtual Machine) environment. The front-end module has a graphical interface that assists the user in CT scan interpretation. The client application runs in any browser that supports JavaScript. The source code of the FAnTom software is available at https://github.com/Center-of-Diagnostics-and-Telemedicine/FAnTom.git.

2.2.3. Clustering model

In our approach, a putative lung nodule is marked with a sphere the center and diameter of which correspond to the center-of-mass and the size of the lesion, respectively. There are two possible scenarios when this strategy can lead to sub-optimal results.

First, the lung nodule usually has a spherical shape (Fig. 2A), but it can be more complex: elongated or consisting of several spherical abnormalities, being distorted by the surrounding tissue (Fig. 2B and C). Description of such lesions with a single sphere would include a large volume that is not part of the nodule. Our approach allows marking a complex-shaped abnormality with a set of overlapping spheres, more accurately approximating its geometry. Whether the spheres correspond to the same nodule is decided on the condition that the distance between two individual sphere centers should be less than the sum of the spheres' radii.

Second, in the double-reading approach, the marks made for the same lesion by different radiologists may not match exactly. According to Revel et al., the inter-reader variability in diameter measurements can reach 20% of the average nodule diameter [39]. Another source of variability is the location of the center-of-mass of the lesion. Finally, the readers may have different interpretations of complex-shaped nodules. Some readers may approximate them with a set of spheres, while others use a single sphere of larger diameter (Fig. 2B). If one reader's sphere contains the center of the other reader's sphere, we refer these marks to the same nodule (Fig. 2A). If two or more neighboring spheres made by one or different readers just overlap, they are combined into a cluster that describes a complex-shaped lesion (Fig. 2Band C).

2.2.4. Image annotation

The readers performed blinded annotation and markup of CT scans using dedicated in-house FAnTom software to indicate the diameter of abnormality and coordinates of the center-of-mass. The FAnTom interface displays the marks made by the radiologist on all the three sections of the DICOM image (transverse, coronal, and sagittal), allowing to adjust the marks on any of them. The interface also contains two fields for specifying the nodule type (solid, part-solid, or ground-glass) and its malignancy. The resulting report was written to a separate JSON file, individual for each reader and each CT scan.

Next, the six reports for a scan were processed to cluster the marks following the above conditions. The clustering results were written to a new JSON file. The combined report for each study contains separate records for all identified clusters. Each cluster consolidates all marks made by the readers, including empty marks for those who did not mention the object. Due to possible disagreements between the readers on whether a particular entity is a lung nodule, up to five records in a cluster can be empty.

Finally, the clustering results were reviewed by the arbitrators. The arbitrators were not able to add new lesions to the JSON file.

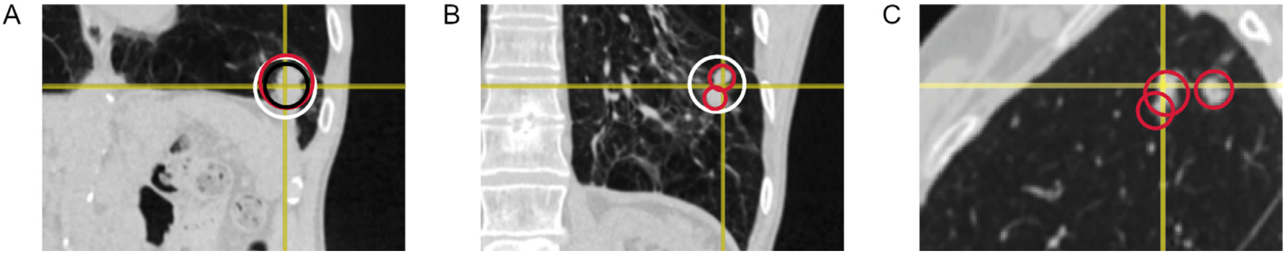

Fig. 2. Diversity of lung nodules' shapes and their markup. Colors of the spheres correspond to the marks made by different radiologists. *A*: markup of a basic nodule; *B*, *C*: variants of the markup of a nodule with a complex shape.

Their task was to inspect each mark and verify it according to the following criteria:

1. The arbitrator's agreement with the presence of a nodule at the site. If the answer was negative, the "rejected" value was assigned to the *decision* key, and no further review was needed. If the arbitrator approved the mark, the nodule type and size adjustment had to be made. If there was a complete agreement, the arbitrator assigned the "confirmed" value to the *decision* key.
2. The arbitrator could disagree with the initial assessment of the lesion type or size. In this case, the *decision* key got the "confirmed partially" value and the explanatory statement in Russian was assigned to the *comment* key (for example, "type mismatch", "incorrect size"). Besides that, in case of uncertainty the arbitrator had the option to use the "doubt" value for the field. The *type* key was intended for the arbitrator's opinion on the lesion type. In addition to the three above-mentioned nodule types, the arbitrators were able to use the "other" type to mark dubious entities. The *proper size* key contained the arbitrator's opinion on the initial size estimate, taking one of two values: "true" or "false".
3. The arbitrator's estimate of nodule malignancy, with two possible values for the *malignancy* key: "true" and "false". The "false" value was assigned to abnormalities containing fat, fibrous tissue, fluid, and other benign and non-nodular lesions.

2.2.5. JSON file structure

Annotation data in the JSON file consisted of three sections:

1. The ***doctors*** section containing the information about the reader's numeric ID and the comment made by the radiologist during the initial annotation stage. The content of comments was not formalized; they could relate to such issues as the presence and type of abnormalities or the initial diagnosis. Therefore, the comments were not translated into English and are given only in Russian. Every reader had a three-digit personal ID; if a radiologist was replaced during the image annotation process, the new reader inherited the ID of the predecessor with an additional "+" symbol.
2. The ***ids*** section containing the list of study identifiers obtained from the DICOM headers: *study ID, accession number*, and *study instance UID*. The section also included information on the patient's age and sex (*age* and *gender* keys). The value for the *age* key had a four-character format with the first three positions describing the numerical value, and the last position specifying the unit type for the age (Y, for years; M, for months; D, for days). The value for the *gender* key was either "F" (female) or "M" (male).
3. The ***nodules*** section with the data on lesions identified by the radiologists. Here resides an array of cluster records, each containing six reader records for a putative nodule. Up to five objects can be empty if the reader did not identify the nodule. Each non-empty object contained the following keys:
   (a) parameters of the sphere encircling the abnormality: *diameter* (mm) and coordinates of the center (*x*- and *y*-coordinates are always in pixels, *z*-coordinate is in mm). We observed a few cases of skipping slices in CT scans due to PACS downloading peculiarities. To avoid the *z* scale shift, rather than the slice number, we used its absolute coordinate in mm recorded in the DICOM attribute Image position (patient), tag # (0020,0032);
   (b) the type of the nodule, with one of three values: "#0S" (solid), "#1PS" (part-solid), and "#2GG" (ground-glass);
   (c) the *expert decision* key containing the results of arbitration.

2.2.6. Reader accuracy and inter-observer agreement

For the accuracy *(Acc)* calculations, cases when, in the arbitrator's opinion, at least one reader correctly identified a lung nodule at the specific site of a CT scan were recognized as true positives (TP). True-negative (TN) results included cases when, in the arbitrator's opinion, at least one reader did not mark an entity that was incorrectly marked by any other reader. For the data analysis, we assumed that the arbitrator's judgment was always correct. The accuracy was calculated as:

$$Acc = \frac{TP + TN}{P + N}, \tag{2}$$

where *P* is the number of correct findings, and *N* is the number of incorrect findings.

The inter-observer variability was analyzed using the percentage agreement metric. The statistical analysis was performed with R 3.6.3 [40] using `dplyr` [41] and `irr` [42] packages.

2.2.7. Database access

The DICOM images and associated JSON files for all 536 cases are available at https://mosmed.ai/en/datasets/ct_lungcancer_500/.

3. Results

The participants providing their data for the dataset were 60% females and 40% males aged 50 to 75 years (62.3 ± 6.2 and 62.4 ± 8.7 for females and males, respectively). For 72 subjects, radiologists did not find any lung abnormalities. The remaining 464 CT scans contained 3151 nodules marked by the readers.

Of these, 1761 (55.8%) nodules were recognized as malignant, and 445 (14.1%) nodules were assigned to the category of benign lesions. There were also 926 (29.4%) abnormalities of non-nodular type. In the remaining 19 cases, the radiologists did not agree on the nodule's malignancy (Supplementary Table S2).

Thirty-one radiologists performed the initial reading. Each of the 15 radiologists from the initial pool was replaced at some point of the study due to refusal or inability to further participate in the project; in one case, the replacement reader was also replaced.

3.1. Accuracy and efficiency of different markup strategies

The nodule size is one of the most important parameters for diagnostic accuracy [43]. According to the Fleischner Society guidelines, the nodule size measurements should be performed in two dimensions, along the short and long axes of the lesion [30]. Therefore, for every marked nodule, reader should specify six parameters: coordinates of the center-of-mass, the two size measurements, and slope angle for one of the axes. There are alternative approaches describing the lung nodule shape and size: using a 3D box, a sphere, or constructing a complete outline of lesion. Each of them requires the reader to define a specific set of parameters (Table 1). We compared the accuracy and labor intensity of these approaches in a study of 10,000 computer-generated 3D lung nodule models of different shapes and sizes. Every model was approximated by a single shape in case of the 3D box, sphere, or ellipsoid approaches.

Nodule markup performed by the maximum transverse diameter ("Sphere" method in Table 1) allowed us to achieve the optimal balance between the accuracy and operational time. In terms of the nodule shape approximation, the approach was slightly inferior to the method of nodule description with an ellipsoid model (d 0.5 ± 0.3 versus 0.6 ± 0.3, respectively), but significantly better than the 3D box approach (d 0.3 ± 0.2). The maximum diameter approach was the fastest of all three of them (Table 1). The outline construction method was characterized by the highest accuracy achieved at the expense of the markup time (Table 1).

According to the results of this experiment, the sphere approach to the markup of DICOM images became the basis of our cluster model for nodule localization. The use of additional spheres for a better approximation of the nodule shape can increase the accuracy of the method while maintaining its efficiency.

On average, each of the 15 radiologists from the initial pool marked and annotated 1, 050 ± 140 abnormalities during the database creation using our method, spending about 12 min per CT scan. Their replacements had a much lower workload, with 110 ± 42 marked lesions per reader. A total of three working weeks were spent on the markup and annotation of CT scans.

3.2. Number of readers and accuracy of interpretations

Our dataset contained 2003 (63.6%) lesions identified by only one out of six radiologists, of which 896 were marked as malignant, 324 as benign, and 783 as non-nodular (Supplementary Table S1). Besides that, there were 41 putatively malignant nodules with no type assigned. The number of abnormalities considered as nodules by more than one radiologist decreased gradually with an increase in the number of experts that marked the lesion (Fig. 3).

In total, there were 242 (7.7%) lesions marked by three readers, 118 (3.7%) lesions identified by four readers, and only for 70 (2.2%) abnormalities all the six readers were unanimous (Fig. 3, Supplementary Table S1). The average agreement for pairs of readers was 60.5 ± 5.3%, ranging from 40.2 to 73.0%. The majority of disagreements (93.0 ± 4.1%) were in regard to the presence of a nodule at the specific site on a CT scan.

There was a significant negative correlation between the accuracy of reader interpretations and the inter-observer agreement, as the number of readers increased from two to six, $r = -.78$, $p < 0.05$ (Table 2). In agreement with Herman and Hessel [35], 85.7%, 11.4%, and 2.9% of false positives were made by one, two, and three out of six readers, respectively. For the false-negatives, the distribution was more even, as 25.8%, 8.1%, 8.1%, 19.3%, and 30.6% of them were made by only one, two, three, four, and five out of six readers, respectively.

Table 1
Comparison of different approaches of nodule markup.

Method (Software)	Number of specified parameters	Specified parameters	Markup time, s	Sørensen–Dice coefficient d
3D box (IKT-SNAP)	6	Coordinates of a center, width, height, and depth	18.8 ± 1.5	0.3 ± 0.2
3D box (3D Slicer)	6	Coordinates of a center, width, height, and depth	18.2 ± 1.1	0.3 ± 0.2
Outline construction (ITK-SNAP)	100–1000	Coordinates of all points used for outline construction	91.1 ± 8.9	1.0 ± 0.0
Outline construction (3D Slicer)	100–1000	Coordinates of all points used for outline construction	90.4 ± 2.4	1.0 ± 0.0
Sphere (FAnTom)	4	Coordinates of the center, maximum transverse diameter	5.9 ± 0.3	0.5 ± 0.3
Ellipsoid (FAnTom)	6	Coordinates of the center, long- and short-axes diameters, angle of one of the axes	18.4 ± 1.2	0.6 ± 0.3

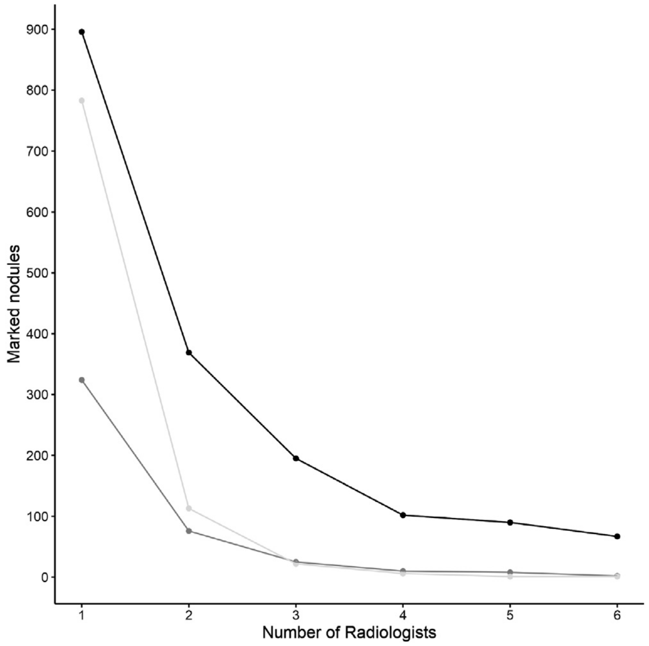

Fig. 3. The number of abnormalities marked by a corresponding number of radiologists. Black: malignant nodules; gray: benign nodules; light-gray: non-nodular lesions.

Table 2
Inter-reader agreement and accuracy.

Number of readers	Agreement, %	Accuracy, %
2	57.0 ± 15.6	79.7 ± 4.9
3	37.1 ± 7.3	89.2 ± 5.1
4	16.5 ± 5.7	93.8 ± 3.6
5	9.8 ± 8.1	97.9 ± 0.1
6	2.2	100.0

3.3. Nodule content in the dataset

For nodule identification, the radiologists were instructed to pay attention to the lesions with sizes ranging from 3 to 30 mm. As a result, the dataset contained only 11 nodules < 3 mm (0.3%) and 42 nodules > 30 mm (1.3%) marked by the readers. For further analysis, we divided our dataset into three subcategories by nodule size according to the recommendations of the Fleischner Society guidelines [30] (Table 3).

There were 72 benign nodules assigned by the experts to the “other” type, and 811 non-malignant abnormalities with no type assigned. Fifteen lesions classified by the radiologists as malignant were also annotated as belonging to the “other” type. All of the 15 nodules were initially recognized as solid or part-solid lesions, but the assessment was not confirmed by an arbitrator. The same is true for the 34 malignant nodules with no type assigned. We believe that these cases require further inspection and plan to address them in the future updates of the dataset.

4. Discussion

The size and quality of a training dataset are the key factors for ML model performance in any application, including medical imaging [44]. Unfortunately, there are no standardized rules and guidelines on how to annotate medical image data properly. Almost every available collection of clinical CT scans has its own information organization design, with its advantages and limitations. Creating new datasets is a time-consuming and challenging task that requires human experts to provide a ground truth annotation. Crowdsourcing performed by unskilled individuals proved itself as a valuable tool for time-efficient large-scale annotations of image databases [24,45], but the accurate interpretation of radiological images requires specialized training and experience [25]. Weak forms of annotations are a compromise between the two approaches aimed at reducing the required annotation efforts [26,27]. We present a cluster model for nodule localization adapted for collaborative and crowdsourced annotation that simplifies the labeling of DICOM images. Using the in-house FAnTom multitenant software based on this model, we have created CTLungCa-500, a mock-up dataset, containing 536 standard-dose chest CT scans of patients with a high risk of lung cancer.

To choose the optimal primary markup method for our model, we conducted a comparative study of the main approaches for describing the geometry of a lung nodule: using a 3D box, a sphere, an ellipsoid, or manually constructing an outline of the lesion. According to the measurements of efficiency and shape approximation accuracy of 10,000 computer-generated models, the sphere approach showed the best balance between these metrics (see Table 1).

Our cluster model is well-suited for the double-reading approach as it can automatically group the marks made by the readers for the same lesion, even if their interpretations differ in terms of shape, size, and location of center-of-mass of an abnormality. Another advantage of clustering is the ability of our model to effectively describe complex-shaped lesions. Tumors, especially advanced ones, are not always spherical; instead, they can have an irregular and heterogeneous shape. Therefore, the classical maximum transverse diameter approach tends to overestimate both tumor diameter and volume [46]. Our model enables the description of a heterogeneous, non-spherical nodule as a collection of overlapping spheres, which minimizes the inaccuracy of shape approximation.

Currently, radiology reports are a prevalent ground truth annotation method [6,8,47], but it is associated with errors in the detection of lung nodules [31,32]. A common practice to minimize the error is to use the double-reading approach with arbitration of initial readings. For example, the creators of the LIDC/IDRI database used subjective assessments of four experienced radiologists revised at the second read phase considering the interpretations of other readers [6]. Each radiologist constructed a full outline of every nodule and provided a detailed description of the lesion. Unfortunately, such a high-quality and high-effort approach is very time-consuming; it took about seven years to create the LIDC/IDRI database [6]. Despite the fact that the results of other readers were disclosed to the radiologists for a final decision, there still was significant variability in identification and classification of lesions. The main focus of the LIDC/IDRI database was on nodules ≥ 3mm; it contains 2,669 lesions of this category identified by at least one reader. Of these, only 928 lesions (35%) were marked by all the four radiologists.

In our model, we suggest a similar ground truth annotation approach, lowering the annotation quality requirements and increasing the number of initial readers to reduce the probability of false-negative errors. This approach allowed us to get different interpretations of DICOM images, but with the side effect of serious disagreement between the readers. The inter-observer agreement for six readers was only 2%, which is far below that of the LIDC/IDRI database. We associate it with the fact that following the instructions and limiting their efforts at five entities per CT scan, the radiologists marked different lesions in cases of multiple pulmonary nodules. Besides that, the volunteer radiologists performed the interpretation of CT scans in a non-controlled environment, usually at the end of the working day or at night, being exhausted. Moreover, the readers used their personal computers with user-specific FAnTom parameters for brightness, contrast, and gamma

Table 3
Dataset nodule content by size and type.

Nodule type	≤ 6 mm				6 mm < nodule size ≤ 8 mm				8 mm < nodule size ≤ 30 mm			
	malignant		non-malignant		malignant		non-malignant		malignant		non-malignant	
	count	size, mm	count	size, mm	count	size, mm	count	size, mm	count	size, mm	count	size, mm
solid	716	4.7 ± 0.7	112	4.7 ± 0.8	209	7.1 ± 0.6	37	7.0 ± 0.5	384	13.8 ± 5.4	35	12.3 ± 4.5
part-solid	118	4.7 ± 0.8	32	4.9 ± 0.9	35	7.0 ± 0.5	6	6.9 ± 0.2	26	12.6 ± 4.5	12	12.2 ± 3.1
ground-glass	56	4.7 ± 0.7	23	5.3 ± 0.8	25	7.1 ± 0.6	10	7.2 ± 0.6	48	13.5 ± 5.0	9	10.8 ± 1.5
other	7	4.3 ± 0.4	68	4.8 ± 0.7	2	7.2 ± 0.3	19	7.1 ± 0.5	6	17.0 ± 5.5	47	14.9 ± 5.6
none	20	4.4 ± 0.7	444	4.3 ± 0.7	2	7.3 ± 0.9	144	7.3 ± 0.5	12	11.6 ± 2.4	223	12.5 ± 4.1

correction. All these factors, inevitable for a crowdsourcing model, can influence the accuracy of markup and annotation of individual readers. However, the combined accuracy increased with the number of readers (see Table 2). Our results demonstrate that a low-effort annotation can be effective when done by multiple experts and validated by an arbitrator. We plan to address the inter-observer agreement in detail in a separate publication.

The essential feature of training and testing datasets for machine learning is a balance between classes. Cases representing non-cancer tissue are just as important for machine learning as examples of specific pathologies. Our dataset contains 926 marked non-nodular entities and 72 CT scans from patients without any lung abnormalities. Of 2,201 lesions that were annotated as pulmonary nodules, 79%, 12%, and 9% of abnormalities belonged to the solid, part-solid, and ground-glass type, respectively. It is consistent with the results of Henschke et al., according to which solid nodules represented 81% of all positive findings on CT scans, and the remaining 19% corresponded to part-solid and ground-glass nodules [48].

There are several limitations and known issues of this dataset in its current state. First, the sample size (536 CT scans) is not large enough for the training of ML algorithms with a high number of parameters that require statistical significance. We plan to add the remaining 3347 cases to the dataset in the following releases. Second, the metadata contain only radiology reports, with no supporting biological or genomics data. Third, there are several label noise instances: 19 nodules were simultaneously marked as malignant and benign (Supplementary Table S2), 15 malignant nodules were assigned to the “other” type, and 41 malignant nodules had no type assigned; all these cases require additional revision.

Our cluster method also has limitations. In its current state, it does not allow to specify some characteristics of a nodule such as spiculation, subtlety, or internal structure. The method is best suited for the double reading approach and might not be optimal for other strategies. Moreover, dedicated software is required to perform the markup and annotation and display the results.

Despite the limitations, the proposed method provides an efficient tool for the collaborative creation of medical image datasets. It is tolerant of differences in interpretations of the shape and size of a nodule by individual radiologists and allows the reader to describe complex-shaped lesions effectively with low effort. The CTLungCa-500 dataset of standard-dose CT images created using the method represents all categories of lung nodules and healthy tissue, making it applicable for training of CAD algorithms, especially those with relatively few parameters.

5. Conclusion

We present a new simplified cluster model for nodule localization, which minimizes the inaccuracy of tumor shape approximation while utilizing the efficient maximum transverse diameter approach. The model is best suited for collaborative and crowdsourced projects using the double reading approach for CT scan interpretations. It automatically groups the marks made by different readers for the same nodule, even if there is some disagreement on the shape and size of the lesion and location of its center-of-mass.

Using the model, we have created the publicly available dataset of standard-dose lung CT images. The dataset consists of 536 cases collected from lung cancer high-risk patients. Every case is provided with six reports on the location, size, and type of lung nodules verified by the arbitrator. The main purpose of the dataset is the training of ML algorithms; therefore, we looked for a balance between cancer and non-cancer tissues. To make the dataset suitable for training complex CAD algorithms, we plan to expand it to 3883 cases in future releases.

Availability of data and materials

The FAnTom software is available at https://github.com/Center-of-Diagnostics-and-Telemedicine/FAnTom.git. The dataset supporting the conclusions of this article is available at https://mosmed.ai/en/datasets/ct_lungcancer_500/.

Declaration of Competing Interest

Authors declare that they have no conflict of interest.

CRediT authorship contribution statement

S.P. Morozov: Project administration. **V.A. Gombolevskiy:** Conceptualization, Methodology, Methodology, Formal analysis, Writing - review & editing. **A.B. Elizarov:** Software. **M.A. Gusev:** Software. **V.P. Novik:** Software. **S.B. Prokudaylo:** Software. **A.S. Bardin:** Data curation. **E.V. Popov:** Data curation. **N.V. Ledikhova:** Conceptualization, Methodology. **V.Y. Chernina:** Investigation. **I.A. Blokhin:** Investigation. **A.E. Nikolaev:** Investigation. **R.V. Reshetnikov:** Methodology, Formal analysis, Writing - review & editing. **A.V. Vladzymyrskyy:** Project administration. **N.S. Kulberg:** Software, Methodology, Formal analysis, Writing - review & editing.

Acknowledgements

The authors thank Olga Korchazhkina, Senior Research Fellow of Federal Research Center “Computer Science and Control” of the Russian Academy of Sciences, for valuable discussions of the project. The authors are also grateful to Ekaterina Korepina and Nikolay Pavlov for their active and valuable participation in creating the web interface. The authors acknowledge Marina Vlasova for proof-reading the manuscript. Ethical approval no. 2 (1-II-2020) was granted by the regional ethics board, Independent Ethics Committee, Moscow Regional Office of the Russian Society of Radiologists and Radiographists. All participants completed and signed a dedicated consent form prior to study participation.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.cmpb.2021.106111

References

[1] C. Fitzmaurice, D. Dicker, A. Pain, et al., The global burden of cancer 2013, JAMA Oncol. 1 (2015) 505–527.

[2] F. Bray, J. Ferlay, I. Soerjomataram, et al., Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J. Clin. 68 (2018) 394–424.

[3] R. Sangha, J. Price, C.A. Butts, Adjuvant therapy in non-small cell lung cancer: current and future directions, Oncologist 15 (2010) 862–872.

[4] H. Lemjabbar-Alaoui, O.U. Hassan, Y.W. Yang, P. Buchanan, Lung cancer: biology and treatment options, Biochim. Biophys. Acta 1856 (2015) 189–210.

[5] D. Aberle, A.M. Adams, C.D. Berg, et al., Reduced lung-cancer mortality with low-dose computed tomographic screening, N. Engl. J. Med. 365 (2011) 395–409.

[6] S.G. Armato, G. McLennan, L. 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, Med. Phys. 38 (2011) 915–931.

[7] A.A.A. Setio, A. Traverso, T. de Bel, et al., Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge, Med. Image Anal. 42 (2017) 1–13.

[8] S.G. Armato, K. Drukker, F. Li, et al., LUNGx challenge for computerized lung nodule classification, J. Med. Imaging (Bellingham) 3 (2016) 044506.

[9] M. Dolejsi, J. Kybic, M. Polovincak, et al., The lung time: annotated lung nodule dataset and nodule detection framework, in: Medical Imaging 2009: Computer-Aided Diagnosis, volume 7260, International Society for Optics and Photonics, p. 72601U.

[10] B.V. Ginneken, S.G. Armato III, B. de Hoop, et al., Comparing and combining algorithms for computer-aided detection of pulmonary nodules in computed tomography scans: the ANODE09 study, Med. Image Anal. 14 (2010) 707–722.

[11] G. Han, X. Liu, F. Han, et al., The LISS a public database of common imaging signs of lung diseases for computer-aided detection and diagnosis research and medical education, IEEE Trans. Biomed. Eng. 62 (2014) 648–656.

[12] National Cancer Institute. National Lung Screening Trial, 2018, (https://www.cancer.gov/types/lung/research/nlst), Accessed: 2020-10-20.

[13] D. Ardila, A.P. Kiraly, S. Bharadwaj, et al., End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography, Nat. Med. 25 (2019) 954–961.

[14] P. Huang, C.T. Lin, Y. Li, et al., Prediction of lung cancer risk at follow-up screening with low-dose CT: a training and validation study of a deep learning method, Lancet Digit. Health 1 (2019) e353–e362.

[15] J.R. Jett, Limitations of screening for lung cancer with low-dose spiral computed tomography, Clin. Cancer Res. 11 (2005) 4988s–4992s.

[16] Y. Qi, X. Cui, M. Han, et al., Radiomics analysis of lung CT image for the early detection of metastases in patients with breast cancer: preliminary findings from a retrospective cohort study, Eur. Radiol. 30 (2020) 4545–4556.

[17] T. Sun, R. Zhang, J. Wang, et al., Computer-aided diagnosis for early-stage lung cancer based on longitudinal and balanced data, PLoS One 8 (2013) e63559.

[18] J.R. Ferreira Junior, M. Koenigkam-Santos, et al., Radiomics-based features for pattern recognition of lung cancer histopathology and metastases, Comput. Methods Prog. Biomed. 159 (2018) 23–30.

[19] C.T. Rueden, J. Schindelin, M.C. Hiner, et al., Imagej2: imagej for the next generation of scientific image data, BMC Bioinform. 18 (2017) 1–26.

[20] P.A. Yushkevich, J. Piven, H.C. Hazlett, et al., User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability, Neuroimage 31 (2006) 1116–1128.

[21] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, et al., 3D slicer as an image computing platform for the quantitative imaging network, Magn. Reson. Imaging 30 (2012) 1323–1341.

[22] K.A. Philbrick, A.D. Weston, Z. Akkus, et al., Ril-contour: a medical imaging dataset annotation tool for and with deep learning, J. Digit. Imaging 32 (2019) 571–581.

[23] M. Nolden, S. Zelzer, A. Seitel, et al., The medical imaging interaction toolkit: challenges and advances, Int. J. Comput. Assist. Radiol. Surg. 8 (2013) 607–620.

[24] E. Heim, T. Roß, A. Seitel, et al., Large-scale medical image annotation with crowd-powered algorithms, J. Med. Imaging 5 (2018) 034002.

[25] V. Cheplygina, A. Perez-Rovira, W. Kuo, et al., Early experiences with crowdsourcing airway annotations in chest CT, Deep Learning and Data Labeling for Medical Applications, Springer, 2016. Pp. 209–218

[26] T. Schlegl, S.M. Waldstein, W.D. Vogl, et al., Predicting semantic descriptions from medical images with convolutional neural networks, in: International Conference on Information Processing in Medical Imaging, Springer, 437–448.

[27] M. Rajchl, L.M. Koch, C. Ledig, et al., Employing weak annotations for medical image analysis problems, arXiv preprint arXiv:1708.06297(2017).

[28] G.V. Rossum, The Python Library Reference, Release 3.8.2, Python Software Foundation, 2020.

[29] R.L. Figueroa, Q. Zeng-Treitler, S. Kandula, L.H. Ngo, Predicting sample size required for classification performance, BMC Med. Inform. Decis. Mak. 12 (2012) 1–10.

[30] H. MacMahon, D.P. Naidich, J.M. Goo, et al., Guidelines for management of incidental pulmonary nodules detected on CT images: from the fleischner society 2017, Radiology 284 (2017) 228–243.

[31] R. Fitzgerald, Error in radiology, Clin. Radiol. 56 (2001) 938–946.

[32] A.D. Ciello, P. Franchi, A. Contegiacomo, et al., Missed lung cancer: when, where, and why? Diagn. Interv. Radiol. 23 (2017) 118–126.

[33] H. Geijer, M. Geijer, Added value of double reading in diagnostic radiology, a systematic review, Insights Imaging 9 (2018) 287–301.

[34] S.J. Hessel, P.G. Herman, R.G. Swensson, Improving performance by multiple interpretations of chest radiographs: effectiveness and cost, Radiology 127 (1978) 589–594.

[35] P.G. Herman, S.J. Hessel, Accuracy and its relationship to experience in the interpretation of chest radiographs, Investig. Radiol. 10 (1975) 62–67.

[36] R. Peters, M. Heuvelmans, S. Brinkhof, P.V. Ooijen, M. Oudkerk, P. de Jong, R. Vliegenthart, et al., Prevalence of pulmonary multi-nodularity in CT lung cancer screening, European Congress of Radiology, 2015.

[37] N.E. Navin, The first five years of single-cell cancer genomics and beyond, Genome Res. 25 (2015) 1499–1507.

[38] N.S. Kulberg, M.A. Gusev, R.V. Reshetnikov, et al., Methodology and tools for creating training samples for artificial intelligence systems for recognizing lung cancer on CT images, Health Care Russ. Fed. 64 (2020) 343–350.

[39] M.-P. Revel, A. Bissery, M. Bienvenu, et al., Are two-dimensional CT measurements of small noncalcified pulmonary nodules reliable? Radiology 231 (2004) 453–458.

[40] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2020.

[41] H. Wickham, R. François, L. Henry, K. Müller, dplyr: A grammar of data manipulation. r package version 1.0.4, 2021.

[42] M. Gamer, J. Lemon, I. Fellows, P. Singh, irr: Various coefficients of interrater reliability and agreement, 2019.

[43] N. Kothary, L. Lock, D.Y. Sze, L.V. Hofmann, Computed tomography-guided percutaneous needle biopsy of pulmonary nodules: impact of nodule size on diagnostic accuracy, Clin. Lung. Cancer 10 (2009) 360–363.

[44] M.D. Kohli, R.M. Summers, J.R. Geis, Medical image data and datasets in the era of machine learning-whitepaper from the 2016 C-MIMI meeting dataset session, J. Digit. Imaging 30 (2017) 392–399.

[45] S. Albarqouni, C. Baur, F. Achilles, et al., Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images, IEEE Trans. Med. Imaging 35 (2016) 1313–1321.

[46] M. Nishino, D.M. Jackman, P.J. DiPiro, et al., Revisiting the relationship between tumour volume and diameter in advanced NSCLC patients: An exercise to maximize the utility of each measure to assess response to therapy, Clin. Radiol. 69 (2014) 841–848.

[47] J. Wallner, I. Mischak, J. Egger, Computed tomography data collection of the complete human mandible and valid clinical ground truth models, Sci. Data 6 (2019) 190003.

[48] C.I. Henschke, D.F. Yankelevitz, R. Mirtcheva, et al., CT screening for lung cancer: frequency and significance of part-solid and nonsolid nodules, AJR Am. J. Roentgenol. 178 (2002) 1053–1057.