Database Acquisition for the Lung Cancer Computer Aided Diagnostic Systems

Anna Meldo
Clinical Research Center of Specialized Types of Medical Care (Oncological)
St.Petersburg, Russia
anna.meldo@yandex.ru

Lev Utkin, Aleksey Lukashin,
Vladimir Muliukha, Vladimir Zaborovsky
Peter the Great St.Petersburg Polytechnic University (SPbPU)
St.Petersburg, Russia
lev.utkin@gmail.com, alexey.lukashin@spbstu.ru,
mulyukha_va@almazovcentre.ru, vlad2tu@yandex.ru

***Abstract*—Most of the used computer aided diagnostic (CAD) systems based on applying the deep learning algorithms are similar from the point of view of data processing stages. The main typical stages are the training data acquisition, pre-processing, segmentation and classification. Homogeneity of a training dataset structure and its completeness are very important for minimizing inaccuracies in the development of the CAD systems. The main difficulties in the medical training data acquisition are concerned with their heterogeneity and incompleteness. Another problem is a lack of a sufficient large amount of data for training deep neural networks which are a basis of the CAD systems. In order to overcome these problems in the lung cancer CAD systems, a new methodology of the dataset acquisition is proposed by using as an example the database called LIRA which has been applied to training the intellectual lung cancer CAD system called by Dr. AIzimov. One of the important peculiarities of the dataset LIRA is the morphological confirmation of diseases. Another peculiarity is taking into account and including "atypical" cases from the point of view of radiographic features. The database development is carried out in the interdisciplinary collaboration of radiologists and data scientists developing the CAD system.**

I. Introduction

The medical database for machine learning is a set of images, samples, examples for training, validating and testing the computer aided diagnostic (CAD) systems which are used nowadays for implementing the intellectual diagnosis and detection of many diseases especially in oncological area. It is impossible to train an intellectual CAD system for diagnostics diseases without using the corresponding databases. Therefore, a qualitative database acquisition plays an extremely important role in developing the efficient CAD systems.

It should be noted that there is a sufficiently large number of databases suitable for the development of CAD systems. But there are some disadvantages of all of them. In many databases, data is badly structured, it does not correspond to medical society recommendations; it is not identical from the point of view of scanning protocols. Besides, often there are not class labels, for example, the class "lung cancer" in the most popular data collections. Therefore, our goal is to propose some principles or elements of a general methodology of the database acquisition in accordance with "doctor's logic", which can be regarded as a way for implementing the explainable artificial intelligence (AI) in CAD systems as an important component of any system based on applying the deep learning approaches.

Lung cancer (LC) is one of diseases which can be diagnosed by means of the intellectual CAD systems. The scientific interest to this disease is caused by the fact that LC is the most common malignancy in all countries and is a main reason of mortality from tumors [1]. Million new cases of LC are approximately registered annually in the world. The majority of cases are detected in industrialized countries (54%) [2]. That is why one of the perspective applications of AI algorithms in oncology is the LC diagnostic.

It should be noted that computed tomography (CT) is a gold standard of the LC diagnostic. This implies that CT images can be regarded as an optimal base for developing the corresponding CAD systems of LC.

Authors of [3] have divided all cases of lung tissues into four groups from the point of view of simplifying data processing:

1) a nodule does not adjacent closely to vessels and other anatomical structures;

2) a nodule locates at the center of the chest and significantly associates with large vessels;

3) a nodule has a subpleural location;

4) a nodule associates with a small part of pleura.

The above division is used for developing most of the CAD systems for the lung tumors diagnostic [4], [5], [6]. But this peculiarity does not correspond to clinical and radiological classifications of LC, so all of the variants of the diseases cannot be taken into account. There is variability of LC from the point of medical classification and visual analysis view. Moreover, most of the CAD systems are trained on datasets containing mainly nodal types of LC, which make only a part of variants of LC. Therefore, we suppose that it is necessary to involve all variants and properties of LC in the CAD systems. That may help to develop the CAD systems which are closer to "doctor's logic". This concept reflects the aspects that guide the radiologist in making a decision.

We propose a quite new database called Lung Images Resource Annotated (LIRA), which incorporates principles and elements of a general methodology of the database acquisition and takes into account several aspects of the

available information about patients, including information about morphological verification of nodules, about age, sex, smoking, etc. of patients. It is shown how to weigh the data in accordance with the available morphological verification, to apply the Bayesian updating procedure or the naïve Bayesian classifier for improving accuracy of decisions.

We try also to show the aspects which explain the principles of CT interpretation by radiologist. They can be a framework that illustrates a basic concept of collecting and structuring medical images for databases. Besides, there are hardly interpretable cases, which may lead to inaccuracy of results of the CAD system implementation, so they have to be taken into account in the process of the database acquisition. Thus, an association of medical and developer's principles to collect and to process medical images is demonstrated.

The proposed elements of the database acquisition on the basis of the LIRA contributes into several areas, including artificial intelligence, e-health, big data, data mining, storage and management.

II. Open databases in oncology

One of the important types of databases in oncology is open databases. Traditionally, they are used to provide the medical community with the corresponding data to successfully solve problems of machine learning and to develop automatic intellectual diagnostic systems or CAD systems. Open databases are mainly publicly available in Internet. First examples of them were obtained in mammalogy and presented sets of mammograms. The most significant of them is the "Digital database for screening mammography" (DSM) and its updated version of CBIS-DDSM (Curated Breast Imaging Subset of DDSM) created in 1997 [7]. This database includes 2620 digitized mammography studies and contains normal, benign and malignant cases with confirmed or approved information about the pathology. Each study includes two images of each breast, patient information (age at the time of the study, ACR (American College of Radiology) breast density rating, fineness rating for abnormalities, description of abnormalities in ACR keywords), image information (scanner, spatial resolution). We find this approach to collect medical data as the most useful for the machine learning and data processing usage because the datasets contain comprehensive information for implementing efficient CAD systems. A fact is that not every open database of medical images includes all required information. Nevertheless, all of them can be used for developing CAD systems and for comparing the systems each other in terms of informativeness measures.

By the moment, there are many collected open databases for various organs and instrumental modalities. The database BRATS 2015 is used for developing CAD systems for the brain cancer diagnostics [8], the database LiTS is for the liver cancer diagnostics [9]. Medical databases can include images of any modalities. For example, the database BUSIS for diagnostics of breast tumors consists of ultrasound images [10]. The most comprehensive sets of kidney images are CPTAC-CCRCC (chaos.grand-challenge.org/Data), which include CT as well as magnetic resonance images (MRI). The content of databases depends on a type of an analyzed pathology and of a modality which is used for the disease diagnostics. For example, the most informative approach for the prostate cancer diagnostics is the MRI, so the prostate cancer database PROMISE2012 includes MRI images [11]. Medical databases are collected not only from radiological images. Photo images and data of dermatoscopy can also be used. For example, a known CAD system for the melanoma, which has been presented in the famous study [12], was learned with database HAM10000 («Human Against Machine with 10000 training images») [13]. The following Internet sites present many medical open databases: www.iccr-cancer.org/datasets, www.nih.gov, healthdata.gov/dataset, portals.broadinstitute.org/cgi-bin/cancer/datasets.cgi.

The idea of collecting the medical images is historically not new. Public database "Japanese Society of Radiological Technology (JSRT) (Japanese society of radiological technology)" was established in 1998 [14]. The JRST database is a source of data for education, training and research. It has been used in many studies to detect nodules on chest x-rays. Totally database includes 247 digitized chest x-rays. These images are divided in subsets with nodular (154 cases: 100 cases are malignant, and 54 cases are benign) and not nodular pathology (93 cases). The database also contains a text file with the following information about each patient: node size (mm), age, sex, final diagnosis, anatomical location of the node, and classification of nodes as malignant or benign. The interesting peculiarity of marking pathology in this database is the determination of the location of every nodule not only by anatomical position, but also using x and y coordinates on the x-ray image (the upper left corner was defined as the origin of x- and y-coordinates). The identification of nodes on images was carried out by three experienced radiologists. The decisions about a type of pathology were based on the consensus of opinions of three thoracic radiologists. The main advantage of this database is that the images cover a wide range of different types of nodules, and they are of the high quality. However, the database does not include information on calcification of nodules.

National Institutes of Health (NIH) Chest X-Ray Dataset (USA) is a publicly available data source for developing methods for automated detection of main abnormalities of the lungs [9]. The database contains 108948 x-rays of 32717 unique patients. Each image contains text labels with one or more keywords (atelectasis, cardiomegaly, effusion, infiltration, mass, nodule, pneumonia, pneumothorax, or normal).

Early Lung Cancer Action Program (ELCAP) is the action for early treatment of lung cancer (USA) presented in its publicly available database of lung images in 2003. The database consists of a set of images of 50 documented low-dose computed tomograms of the lungs for detection obtained in one breath delay with a slice thickness of 1.25 mm. Location of the nodules detected by the radiologist is also provided in [15].

LISS is a publicly available database of common visual signs of lung diseases for automated detection and diagnosis, research, and medical education. It contains 271 chest CT and

the corresponding annotations of the radiologist. The main peculiarity of the LISS database is that it is developed from a new point of view of CT images. The condition of division into subsets is a sign of lung diseases instead of the commonly considered nodules in the lungs. 677 abnormal areas are divided into categories of the common lung disease characteristics in the computed tomography images: calcification, cavities, spicules, lobulation, pleural alteration, bronchial mucus plugs, obstructive pneumonia, air bronchogram [16].

One of the biggest collections called by TCIA (The Cancer Imaging Archive) of medical images is described in the paper [17]. TCIA is a service that contains a large archive of medical images. One of the missions of TCIA is to encourage and to support the open oncological science communities. This content is posting collections of images and providing the search of metadata in the repository. The data are grouped into collections by cancer type, by image modality (MRI, CT, PET (Positron Emission Tomography), x-rays), and by the studied organs (lungs, brain, etc.). The collections can contain images of different types. The main file format used by TCIA to store images is DICOM (Digital Imaging and Communication in Medicine). Additional data such as patient outcomes, treatment, genomics, pathology, and expert analyses may also be provided with the images. Regarding lung cancer, the archive contains 21 collections, including datasets divided by the histological types. For example, the collections "Clinical Proteomic Tumor Analysis Consortium Lung Adenocarcinoma (CPTAC-LUAD)" and "The Cancer Genome Atlas Lung Adenocarcinoma (TCGA-LUAD)" include a series of pictures with adenocarcinoma of 19 and 69 patients, respectively. The collection "The Cancer Genome Atlas Lung Squamous Cell Carcinoma (TCGA-LUSC)" is a dataset from 37 patients with squamous cell lung cancer; 7 collections contain images of the non-small cell lung cancer, etc.

The most famous and notable medical databases of chest images are "The Lung Image Database Consortium" (LIDC) and "Image Database Resource Initiative" (IDRI). LIDC was organized by National Cancer Institute (USA) in 2001. Five medical institutions participated in the data collection. Then in 2004, they were joined by additionally two academic centers and 8 private companies. The idea of this consortium was accompanied by active participation of the Food and Drug Administration (FDA), which determined the success of the results. Currently LIDC-IDRI is used as an international resource for developing, learning, and testing CAD systems of pulmonary nodules. It is based on chest CT images. The volume of this bank of medical images is 1018 computed tomograms from 1000 patients associated with annotations of radiologists. The pathology was divided in 3 categories which were interpreted by four independent radiologists: "nodule > or = 3 mm" is defined as any lesion that is considered as the nodule with the largest dimension 3-30 mm regardless of the intended histology; "nodule < 3 mm" is defined as a nodule smaller than 3mm, which is not clearly benign; "non-nodule > or =3 mm" is defined as any lung tissue which is not determined as a nodule. The interpretation process was performed in two phases: blinded-read phase, when each radiologist independently reviewed each CT scan and marked lesions belonging to one of the three categories; unblinded-read phase, when each radiologist independently reviewed their own marks along with the anonymized marks of the three other radiologists to render a final opinion. A goal of this process was to identify as completely as possible all lung nodules in each CT scan without requiring forced consensus [18].

The LUNA-16 (LUng Nodule Analysis) database contains 888 computed tomograms with nodules ≥3 mm. It is the subset of LIDC-IDRI with the exception of cases with a slice thickness more than 2.5 mm. The data validation was also carried out using opinions of four experienced radiologists, however, without morphological verification.

There are also closed (non-public) databases which include less information; their sets are mainly used training and testing CAD systems with some specific targets. In contrast to open databases, groups of patients in closed databases are defined more accurately, but they are not available for free use.

III. DATABASE LIRA

Database LIRA is a replenish dataset which contains CT images of morphologically confirmed cases of lung cancer and other diseases which look similar to LC on the CT images. In order to create dataset for developing a CAD system with a possibility of differential diagnosis, we take into account the "biological behavior" of LC. The structure of LIRA is proposed by radiologists from St. Petersburg Clinical Research Oncological Center and data scientists from St. Petersburg Peter the Great Polytechnic University. LIRA consists of more than 450 fully anonymized chest CT studies with a slice thickness of no more than 2.5 mm. The RadiAnt DICOM Viewer has been chosen for the modifying and dealing with the dataset. The prepared dataset was archived on the server of the Oncological Center before the morphological diagnostics. Then the data was anonymized with the special software DicomCleaner™, i.e., any identifications of CT-scans were changed by means renaming them with a special code in according with the internal hospital register.

A. Lung cancer signs as a base for collection data

A process of the CT image interpretation consists of the structural (anatomical) and densitometric analyses. The structural analysis involves the image evaluation and its "comparison with a norm representation". This type of analysis is applicable to anatomical objects that are distinguished by means of the "human eye" (mediastinal structures, bronchi, vessels, bones). This aspect is important for understanding the clinic-anatomical classification of LC formulated by the medical community, which is based on the anatomical structure of the lung [19]. From the point of view of localization or the level of lesion in the bronchial tree, there are the central cancer (growing from the trachea, the main lobar and the mouths of segmental bronchi), and the peripheral cancer as a tumor that grows from the smaller bronchi, bronchioles. The Pancost cancer is selected in a particular group; it involves not only the lung, but also structures of the neck and the supraclavicular area, destroying the bones.

The division of the bronchial tree, which is not distinguishable by the eye, cannot be subjected to the structural analysis. In this case, the densitometric analysis is used. The CT radiodensity of a lung tissue as well as pathological tissues is evaluated in accordance with the Hounsfield scale which is regarded as a standard quantitative scale for radiodensity. A lung tissue filled by air has the lowest density (-1000 HU). In accordance with a classification of LC, which has been developed by the international community of studying of LC phenotypes, LC may be different. Adenocarcinoma may have different signs which depend on morphological types. The most often observed type is the invasive adenocarcinoma. Typical issues of the invasive adenocarcinoma are nodular shapes, fuzzy contours with radial spicules, solid, not homogenous structures (see Fig. 1). Solid structures from the point of the CT analysis view are totally dense compact plots which are close to a spherical shape.

The lepidic adenocarcinoma can be presented on CT scans as the so-called ground glass opacity area (see Fig. 2).

The squamous cell carcinoma is a lung tumor which grows from cells lining the inner wall of bronchus. It has another features on the CT scans. For this type of cancer, we see a quite different CT image where, in most cases, the air cavity of the decay is seen in the tumor structure (see Fig. 3).

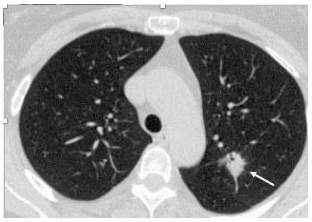


Fig. 1. A chest computed tomogram. There is the adenocarcinoma in the left lung (depicted by the arrow) with typical issues: a nodular shape, a fuzzy contour with radial spicules, a solid and not homogenous structure.

Fig. 2. A chest computed tomogram. There is the lepidic adenocarcinoma in the left lung (arrow), its structure is presented as a ground glass opacity area.

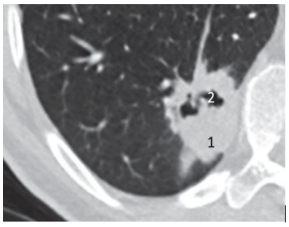


Fig. 3. The squamous cell carcinoma in the right lung (1), its contour is fuzzy and spiculed, there is an air cavity in the lesion (2).

It should be noted that the forms of lung cancer presented in Figs. 1-3 are typical and have pathognomonic features, that is, their diagnosis is beyond doubt. Cases with typical pathognomonic CT-portrait are referred to the group “typical lung cancer”. It is a subset of cases which are interpreted as a cancer by 3 radiologists and are morphologically confirmed as a cancer as well.

At the same time, it turns out that analysis of the clinical material in the St.Petersburg Clinical Research Center of Specialized Types of Medical Care (Oncological) shows that only 65% of LCs have a typical picture, while 26% of CT images correspond to different diseases which require an additional differential diagnostic criteria, 9% of lung cancer cases are extremely difficult to recognize by means of CT due to an atypical visualization image [20].

B. Atypical cases in the database

We analyzed CT-portrait and morphological results of 124 patients. Under atypical cases, we mean those that are lung cancer morphologically but do not have classical signs on CT scans. The examples are presented on Figs. 4 and 5.

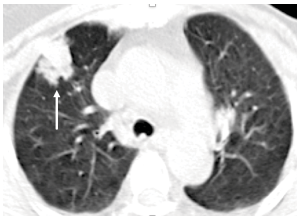


Fig. 4. The chest CT scan shows the pathological lesion which does not have signs of lung cancer (arrow). The preliminary diagnosis was pneumonia, but finally the morphological test gives the LC result.

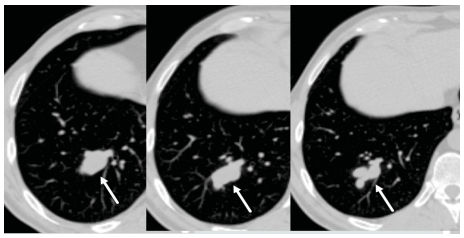


Fig. 5. The pathology lesion in the right lung has not signs of lung cancer (arrows), it looks like the retention cyst, but finally the morphological result is lung cancer.

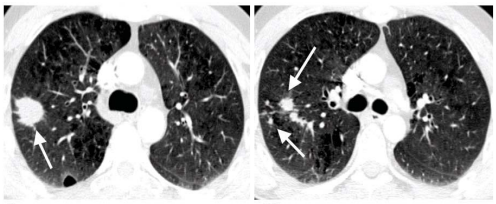
Other false negative cases are aspergilloma – 1, pneumonia – 5, tuberculosis – 2, nontuberculous mycobacteriosis – 1, sarcoidosis – 1, metastases of other malignant tumors – 3. Cases with an atypical CT-portrait are referred to the group “atypical lung cancer”. It is the subset of cases which are interpreted as a “not cancer” by at least 1 of 3 radiologists and are morphologically confirmed as a cancer as well.

The subset “atypical lung cancer” is useful for one-short learning algorithm which is used for training the proposed CAD system.

C. The “not lung cancer” group in the database

The third group in database LIRA is “not lung cancer”. It includes benign tumors, cases which look similar to LC on CT scans, but morphologically they are not LC. They can be regarded as true negative and false positive results of the preliminary conclusion of radiologists and the normal chest CTs.

From the point of view of CT differential diagnostics, tuberculosis is the most common and closest to LC for which the corresponding example is shown in Fig. 6A. Bronchogenic drop-out lesions can be differential feature on CT scans (see Fig. 6B), but they are not always visible.



A. B.

Fig. 6. A: The chest computer tomogram shows tuberculose focus in the right lung which looks like an adenocarcinoma (arrow). B: In addition to main focus, we found bronchogenic drop-out lesions (arrows) – the pathognomonic feature of tuberculosis.

Among false-positives, the most rare and interesting cases are the primary lung lymphoma (see Fig. 7) and the lung abscess (see Fig. 8). The most often and difficult cases for the differential diagnosis are metastases of tumors of other locations. Those cases are included into the subset "not lung cancer" as well.

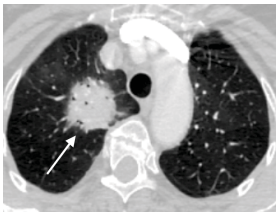


Fig. 7. The chest CT scan shows the focus of the primary lymphoma in the right lung (arrow).

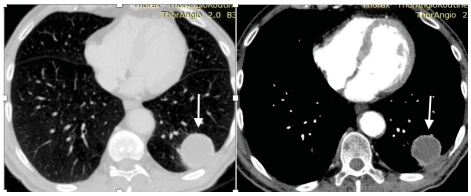


Fig. 8. The chest CT scans show the abscess in the left lung (arrows).

Benign formations are included in the dataset for training the CAD system in the differential diagnosis of LC. Usually the benign lesions have pathognomonic features such as a smooth clear contour and calcium inclusions in the structure. Not all benign formations are morphologically verified. Some of them are confirmed by the follow-up results. No increase of size during 12 months indicates that the corresponding lesion is a benign. 60 chest CTs without any pathology are included into the subset "not lung cancer" too. They are required for testing the CAD system and for calculating the specificity and sensitivity measures characterizing the system.

There are 70 cases in the database LIRA, which are not morphologically verified. They are CTs of patients who refused surgery or the biopsy was impossible due to contraindications. That data was included in the training sample after conclusion of 3 radiologists with the mark "not confirmed". In order to take into account unverified data, we apply an approach which is based on the fact that the weight of each feature vector representing a training instance in the loss function is replaced with a new weight. In particular, if the corresponding instance is from the verified part of the training set, then its weight becomes equal to

$$w_1 = 1 / (N\,[1\text{-}p + q{\cdot}p]),$$

where w is the replaced weight, N is the number of cases in the training set, p is the fraction of the unverified cases, q is the probability of the doctor's correct decisions.

If the instance is unverified, then its weight becomes equal to

$$w_2 = q / (N\,[1\text{-}p + q{\cdot}p]).$$

To calculate q, we analyze the dataset analysis results provided by five independent radiologists who evaluated a part of the verified samples in accordance with the criterion "cancer" / "not cancer".

Summary: the database LIRA contains more than 450 fully anonymized chest CT examples with a slice thickness of no more than 2.5 mm. It consists of 3 subsets in accordance with the primary conclusions of 3 radiologists and the final morphological confirmation of the disease. "typical lung cancer" is a subset of cases which are interpreted as a cancer by 3 radiologists and are morphologically confirmed as a cancer as well; "not typical lung cancer" is a subset of cases which are similar with other diseases on CT scans, and are interpreted as "not cancer" at least by 1 from 3 radiologists and morphologically identified as lung cancer; "not lung cancer" is a subset of cases which are not lung cancer on the basis of morphology or by follow-up normal chest CTs. The size of the nodule is not taken into account. LIRA differs from other open datasets by assigning the class labels "lung cancer". The morphological confirmation of the pathology is an advantage for minimizing false positives in developing machine learning algorithms.

IV. Improvement of the Differential Diagnosis Algorithm by New Class Labels

In order to improve the machine learning algorithms for the differential diagnosis of lung cancer, the dataset LIRA is supplemented by information which represents every case in accordance with the international classification of diseases (ICD).

We take a personal label to each case from LIRA (see Table l). From the point of the differential diagnosis view, this approach to assign class labels is more exact. This can be viewed as a transformation of LIRA into a new format called by LIRA2019.

A main problem that arises by implementing the CAD system with the differential diagnosis of LC is possible small numbers of atypical cases of LC and cases which look similar to cancer in the dataset. This problem can be solved by applying an idea of the few-short learning [21] which is a framework for dealing with the small number of training examples in some classes. There are several methods for solving the few-short learning problem. One of them is to use

the so-called Siamese neural network [22], [23]. The Siamese neural network is composed of two identical neural networks with shared parameters. It aims to compare pairs of examples and to make decision about semantic similarity or dissimilarity of examples in every pair.

TABLE I. CLASS LABELS FOR LIRA2019

Class label	Diagnosis	Notice
D13	Benign tumor	Hamartoma, fibroma, etc. Allowed the confirmation by follow up
D02	Carcinoma in situ	Lepidic carcinoma, morphological confirmation is required
C34	Malignant carcinoma	Morphological confirmation is required
J18	Pneumonia	Allowed the confirmation by follow up
J85.2	Lung abscess	Morphological confirmation is required
I26	Pulmonary embolism	Allowed the confirmation by follow up
D38	Other benign formations	Perifissural nodules, small ground glass opacity focuses without increasing of size, calcifications, local fibrosis.
Mts	Metastases of tumors of another locations	Morphological confirmation is required
A15.1	Tuberculosis confirmed by culture only	For tuberculosis-oriented centers and hospitals
A15.2	Tuberculosis confirmed histologically	There is no culture, but diagnosis is confirmed after the operation
A15.3	Tuberculosis confirmed by unspecified means	There is medical statement of tuberculosis from the tuberculosis-oriented centers

To replenish the database LIRA2019 by partnering hospitals a cloud service was implemented.

V. ADDITIONAL CLINICAL INFORMATION ABOUT PATIENTS

It is well-known that clinical information about a patient is very important for diagnostics of the diseases, because some of them may have similar radiological portrait. For example, some clinical situations may contradict to their visualization. It can be seen from Fig. 9 that there is a chest CT scan with the formation which has features of LC. But symptoms contradict with the LC diagnosis because of the acute pain in the right lower part of the chest after physical activity. So, the diagnosis is the pulmonary embolism, infarct in lower lobe of the right lung.

This example illustrates the necessity of additional clinical information for improving the CAD system development. One of the problems, which has to be solved, is how to use additional features in the form of additional data about patients, including age, sex, smoking, etc. If we would concatenate these features to images and jointly classify the concatenated vector, then we may meet some difficulty related to masking this important information by the large number of features characterizing only nodules. Therefore, we propose to apply the Bayesian updating procedure.

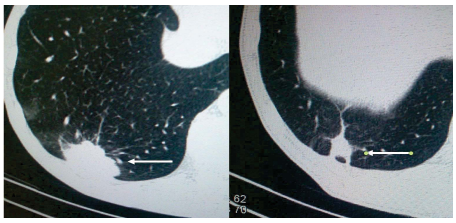


Fig. 9. CT scans of a patient with the pulmonary embolism, infarct in the lower lobe of right lung. On the left CT scan, the formation in the lower lobe of right lung has a signs of lung cancer (the fuzzy contour with radial spicules, the solid structure). On the left CT scan, the formation is decreased without any treatment.

First, we have to define prior probabilities that the investigated nodule is malignant or benign. These probabilities are taken from the output of the CAD system, i.e., the prior probability of the malignant nodule X is $P_{\text{malign}}(X)$. Hence, the probability of benign is $P_{\text{benign}}(X) = 1 - P_{\text{malign}}(X)$. Second, we have to find conditional probabilities $P(x_i = a \mid Y)$, where $Y \in \{\text{malignant, benign}\}$, x_i is an additional feature with index i, i=1,…,m. Here m is the number of features. The conditional probability $P(x_i = a \mid Y = d)$ can be estimated by using the generalized Laplace estimate to avoid the situation when $\#x_i(d) = 0$ as follows:

$$P(x_i = a \mid Y = d) = \frac{\#x_i(a,d) + 1}{\#x(d) + s_i}.$$

Here $\#x_i(a,d)$ is the number of examples in the training set from class d where the i-th feature has the value a; $\#x(d)$ is the number of examples in class d; s_i is the number of possible values for x_i. For example, if 100 patients have cancer and 80 people among these 100 patients are smoking, then we can write that $s_i = 2$ (smoking, non-smoking), $\#x_i(\text{smoking}, \text{malignant}) = 80$, $\#x_i(\text{malignant}) = 100$, and

$$P(x_i = \text{smoking} \mid Y = \text{malignant}) = \frac{80 + 1}{100 + 2}.$$

By having the conditional probabilities $P(x_i = a \mid Y = d)$, prior probabilities of malignant or benign for a new patient can be updated and the corresponding posterior probabilities are computed as follows:

$$P^*_{\text{malign}}(X) = P(Y = \text{malignant} \mid X)$$

$$= C \cdot P_{\text{malign}}(X) \cdot \prod_{i=1}^{m_{\text{actual}}} P(x_i = a_i \mid Y = \text{malignant}).$$

The posterior probability of benign is defined in the same way:

$$P^*_{\text{benign}}(X) = P(Y = \text{benign} \mid X)$$

$$= C \cdot P_{\text{benign}}(X) \cdot \prod_{i=1}^{m_{\text{actual}}} P(x_i = a_i \mid Y = \text{benign}).$$

Here C is the aspect ratio which is determined from the condition $P^*_{\text{malign}}(X) + P^*_{\text{benign}}(X) = 1$; $m_{\text{actual}} \le m$ is an actual number of known features about the patient. The introduction of m_{actual} stems from the fact that some information about an investigated patient may be unknown. Therefore, we use only features which are available for the patient. In particular, if there is no additional information about a patient, then $P^*_{\text{malign}}(X) = P_{\text{malign}}(X)$ and $P^*_{\text{benign}}(X) = P_{\text{benign}}(X)$, i.e., prior and posterior probabilities coincide.

For collecting information in accordance with this approach, we start to complement the database LIRA2019 by the full anamnestic, clinical, genomic information. A new modification of the dataset will be implemented in 2020 with adding materials of St.Petersburg Clinical Research Center of Specialized Types of Medical Care (Oncological) and hospitals being partners of the Center.

VI. Conclusion

Authors think that there are two conditions for successful using the medical CAD systems. The first one is an explainability of the AI system results, which allows doctors to understand how the system makes decision. The second one is a certain correspondence of approaches for developing the CAD systems with the clinical and radiological classifications. This is a pledge of acceptance of the developed CAD systems by the medical community. The collaboration between radiologists and developers of the CAD systems may be useful for interdisciplinary understanding of the system application goals. On the one hand, collecting of specialized datasets increases the range of possible functional tasks of a radiologist. On the other hand, the experience of the radiologist takes part in machine learning algorithms which should be implemented to be closer as possible to doctor's logic.

Finally, the main proposed idea of the methodology of creating the medical databases can be formulated as follows: structuring data, their homogenization, verification of diseases, inclusion of the "atypical" cases and cases which look similar to a studying disease.

The dataset LIRA is a unique dataset which consists of the morphologically confirmed LCs and pathologies looked similar to the LC. It contains class labels for each nodule to develop the differential diagnostic intellectual algorithm. LIRA2019 is a modified dataset which is improved by adding new class labels in accordance with the international classification of diseases (ICD).

The necessity to include additional anamnestic and clinical information is a reason for developing a new database LIRA2020 in near future.

Acknowledgment

The reported study was funded by RFBR, project number 19-29-01004.

The results of the work were obtained using computational resources of Peter the Great Saint-Petersburg Polytechnic University Supercomputing Center (www.spbstu.ru) which is registered as a center of collective usage (http://ckp-rf.ru/ckp/500675/).

References

[1] D.G. Zaridze "Epidemiology and prevention of cancer", *Voprosy oncologii*, no. 9, 2001, pp. 6-14, (in Rus).

[2] Cancer incidence, mortality and prevalence worldwide in 2008. Int. Agency for Research on Cancer (IARC). Web: http://globocan.iarc.fr.

[3] W.J. Kostis, A.P Reeves., D.F. Yankelevitz, C.I. Henschke "Three-dimensional segmentation and growth-rate estimation of small pulmonary nodules in helical CT images". *IEEE Transactions on Medical Imaging*, vol. 22(10), 2003, pp. 1259-1274.

[4] K. Doi "Current status and future potential of computer-aided diagnosis in medical imaging", *The British Journal of Radiology*, vol.78, 2005, pp.3-19.

[5] M. Firmino, A.H Morais., R.M. Mendoca., et al. "Computer-aided detection system for lung cancer in computed tomography scans: review and future prospects", *Biomedical engineering online*, vol. 13(1), 2014, pp. 41.

[6] M.Z. Rehman., M. Javaid, S.I.A. Shah et al. "An appraisal of nodules detection techniques for lung cancer in CT images". *Biomedical Signal Processing and Control*, vol.41, 2018, pp.140-151.

[7] F. Prior et al., "The public cancer radiology imaging collections of The Cancer Imaging", *Archive. Sci. Data,* vol.4, Jan.2017, p.124, doi: 10.1038/sdata.2017.124 (2017).

[8] B.H. Menze, A. Jakab et al. "The multimodal brain tumor image segmentation benchmark (BRATS)", *IEEE Trans. Med. Imaging*, vol.34, 2015, pp 1993– 2024.

[9] P. Bilic., P.F. Christ et. al. "The liver tumor segmentation benchmark (LiTS)", *arXiv*: 1901.04056, Jan 2019

[10] M. Xian, Y. Zhang, H.D. Cheng, F. Xu, et al. "Benchmark for breast ultrasound image segmentation (BUSIS)", *arXiv:* 1801.03182, Jan 2018.

[11] G. Litjens et al. "Evaluation of prostate segmentation algorithms for MRI: The PROMISE12 challenge", *MedIA*, vol.18(2), 2014, pp 359–373.

[12] H.A. Haenssle, C. Fink, R. Schneiderbauer., et al. "Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists", *Annals of Oncology*, 28 May 2018, pp 1 – 7.

[13] P. Tschandl, C. Rosendahl, H. Kittler "The HAM10000 Dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions", *arXiv*,1803.10417, Mar 2018.

[14] JSRT Database [Japanese Society of Radiological Technology Web]: 2004; http://db.jsrt.or.jp/eng.php.

[15] ELCAP Public Lung Image Database Web: http://www.via.cornell.edu/lungdb.html.

[16] G. Han, X. Liu, F. Han, I. Santika, Y. Zhao, X. Zhao, C. Zhouet, "The LISS—A public database of common imaging signs of lung diseases for computer-aided detection and diagnosis research and medical education", *IEEE Trans. Biomedical Engineering*, vol.62(2), Feb.2015, pp.648-656.

[17] K. Clark, B. Vendt, K. Smith et al. "The cancer imaging archive (TCIA): Maintaining and operating a public information repository", *Journal of Digital Imaging*, vol. 26(6), Dec. 2013, pp. 1045-1057.

[18] S.G. Armato 3rd, 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*., vol.38(2), Feb. 2011, pp. 915-931, doi: 10.1118/1.3528204.

[19] M. Prokop, M. Galanski, *Spiral and multislice computed tomography of the body*. Stuttgart-New York: Thieme, 2011.

[20] A.A. Meldo., L.V. Utkin "A computer-aided system for differential diagnosis of lung diseases", *Intelligent Data Processing: Theory and Applications. Book of abstracts of the 12th International Conference,* Moscow, Russia – Gaeta, Italy, 2018, Moscow: TORUS PRESS. - 2018 – P. 35

[21] L. Fei-Fei, R. Fergus, and P. Perona, "One-shot learning of object categories," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 594–611, 2006.

[22] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recog- nition (CVPR'05)*, vol. 1. IEEE, 2005, pp. 539–546.

[23] G. Koch, R. Zemel, and R. Salakhutdinov, "Siamese neural networks for one-shot image recognition," in *Proceedings of the 32nd International Conference on Machine Learning*, vol. 37, Lille, France, 2015, pp. 1–8.