A new approach to differential lung diagnosis with CT scans based on the Siamese neural network

A A Meldo$^{1,2}$ and L V Utkin$^2$

$^1$ St.Petersburg Clinical Research Center of Specialized Types of Medical Care (Oncological), St.Petersburg, Russia
$^2$ Peter the Great St.Petersburg Polytechnic University, St.Petersburg, Russia
E-mail: lev.utkin@gmail.com

Abstract. A lot of computer-aided diagnosis systems for lung cancer detection have been developed in the last years, but most of them may be not effective when we deal with the differential diagnosis. Only a part of patients with lung cancer have a typical nodular cancer. Some patients have the extremely difficultly recognized lung cancer due to its atypical visualization. A new approach is proposed for differential lung diagnosis and for solving the problem of classifying all types of lung cancer. It is based on the following ideas. First, we apply the length chord and Hounsfield unit value histograms characterizing the inner structure of a tissue and its surrounding as a new feature representation of the tissue. Second, we collect a special dataset which contains a lot of atypical cases of cancer and non-cancer tissues. Third, we propose to use histograms as well as suspicious lung object images for training and using two Siamese neural networks which can be viewed as a key element of the proposed approach and allow us to implement some elements of explainable artificial intelligence systems.

1. Introduction
Lung cancer (LC) is the most common malignancy tumor in many countries and the main reason of mortality from malignancy tumors [1]. However, LC can be treated effectively in many cases when it is diagnosed at an early stage, but its diagnostic is rather difficult at early stages. A lot of computer-aided-diagnosis (CAD) systems based on using the artificial intelligence (AI) approach have been developed in order to assist radiologists detecting the pulmonary nodules automatically at early stages [2]. As the most sensitive imaging modality, computed tomography (CT) has strong competitive advantages, such as rapid acquisition and being cost effective and widely available [3]. Therefore, many CAD systems use CT scans as a basis for analyzing LC. Comprehensive reviews of the LC CAD systems architectures can be found in review papers [4, 5, 6, 7].

Technically, many CAD systems consist of two parts: the computer-aided detection subsystem concentrated on segmentation of lung nodules, and the computer-aided diagnosis subsystem which focuses on classification of the suspicious lung lesions. We consider the second subsystem assuming that the first one is perfectly performed.

A detailed analysis of various CAD systems have shown that they are trained on datasets containing mainly nodal types of LC. This peculiarity does not correspond to clinical and radiological classifications of LC, so all cases of the lung disease cannot be taken into account. Therefore, we aim to develop a CAD system which takes into account atypical cases of the lung
disease. Moreover, we aim to develop a system which allows us to explain its diagnostic results. In other words, the proposed CAD system should be regarded as an explainable AI system.

The main difficulty of considering the atypical cases of the lung disease from the point of view of the LC nodal types, including the peripheral lung cancer (squamous cell carcinoma), metastasis, tuberculosis looking like a cancer on CT images, lymphoma, hamartomas, lepidic adenocarcinoma, etc. [8], which do not have nodular structure. These atypical cases do not occur frequently in practice and may be present in training datasets as one exemplar or a few examples. As a result, many CAD systems cannot detect some atypical cases. This situation is similar to the well-known problem of one-shot or few-shot learning [9]. One of the most efficient tools for one-shot learning is the Siamese neural network (SNN) [10], which consists of two identical neural subnets with shared parameters. The SNN does not solve the classification task, but it compares two objects and decides whether these two objects are semantically similar or dissimilar. The SNN may significantly simplify the problem of differential lung diagnosis by learning on the dataset of atypical cases and by exploiting on pairs of objects such that one of them is a new investigated object and another object is taken from the dataset. This is the main idea to implement the differential lung diagnosis. Moreover, according to [11], an important condition of the efficient CAD system is that a doctor must have a possibility to understand how and why a machine decision has been made. The proposed incorporating the SNN provides this possibility, i.e., we get some elements of an explainable AI system.

At the same time, we can use the SNN in two ways. The first way is when 3D images of segmented suspicious objects are fed to inputs of the SNN subnets. In this case, we have to implement the SNN as a pair of 3D convolutional neural networks (CNN). The second way is to use some feature representation of the segmented suspicious objects of the low dimension in order to implement the SNN consisting of two shallow or rather simple neural networks. Therefore, the second important idea used in the proposed CAD system is to represent images of the suspicious objects in the form of histograms which can be viewed as a representation of 3D images. This can be done by using the chord method [12] and its modifications oriented towards a unique feature representation of objects.

In order to train and to use the SNN, it is necessary to have a dataset of atypical cases as well as standard nodules. The dataset collection is the third idea underlying the proposed approach to constructing the CAD system.

2. Chord histograms for nodule feature representation

One of the ways for the low-dimensional feature representing of a suspicious object is the chord length histogram based on the chord method [12]. Many points are randomly selected at the object surface, and their pairs produce a set of chords. The histogram of the chord lengths represent the shape information about the nodule, and histograms corresponding to different types (diseases) of objects are different. This implies that a neural network can be trained by using examples of various object in the form of the chord length histograms.

Since the chord length histogram represent the shape of the object, then it makes sense to consider its inner structure. If we measure the radiodensity (Hounsfield unit) at points randomly selected on every chord, then we can construct the histogram of the radiodensity values inside the object.

In order to take into account the object surrounding, it is proposed to put the object into a virtual cube. Perpendiculars are drawn from random points on every face of the cube to the surface of the object. The obtained segments can be regarded as outer chords which produce the outer chord length histogram. In addition, we produce the radiodensity value outer histogram by taking random points on the outer chords.

In sum, we have four histograms which represent the object image. New histograms for a more adequate representation of the objects can be additionally constructed, for example, an
interesting histogram which shows changes of the radiodensity values inside the object. However, our experiments show that four aforementioned histograms give outperforming results.

3. Complex tumor structures and a dataset
Below we consider only a few examples of atypical cases which may be observed in the lung CT images. All cases can be divided into atypical LC visualizations and diseases which are similar to LC in CT images. First of all, we have to point out the metastases of not lung located tumors and tuberculosis, which may look like a cancer on CT images. Another example is the lung adenocarcinoma (a peripheral tumor that develops from glandular (mucus-producing) cells). The corresponding LC growing from cells lining the inner wall of the bronchus is called the squamous cell carcinoma. For this type of LC, there may be a quite different CT image with the air cavity of the decay available in the tumor structure, which characterizes the “biological behavior” of this tumor. The next example illustrating difficulties of the CT image interpretation for a radiologist is the malignant lymphoma (hematologic disease of the lymphatic system).

Another example is a tumor having inclusions of calcium, which is typical for hamartomas and is not typical for LC. The tumor may be recognized as LC by means of the morphological investigation. This example again illustrates the fact that improvement of CAD systems is required for detecting LC. The last example is LC from the alveolar walls, which is called the lepidic adenocarcinoma. It doesn’t have a nodular structure on CT scans, its image is similar to the pneumonia.

These few examples show that different diseases may have similar objects in CT images including atypical ones. In order to cope with these cases, we have to develop a CAD system which could recognize them. We propose to apply the Siamese neural network which is trained on these examples collected in a special dataset of atypical lung objects. The dataset includes segmented labelled tumors that are divided into three groups: “typical” peripheral LC, “atypical” LC, “not cancer” (cases that can be interpreted as LC with respect to similar patterns in CT images). The dataset contains only verified tissues confirmed by surgery results or by histological investigation.

4. Siamese neural network as a tissue classification model
Approaches associated with the data representation in the form of a set of histograms despite their effectiveness cannot totally cover the atypical cases that have been considered above. However, the main difficulty of training many CAD systems is that the training set contains only a few corresponding examples. This makes an application of the available systems problematic in real medical practice. The situation of unique cases is similar to the one-shot learning for which one of the classification algorithms is the SNN [10]. Therefore, we propose an approach using the SNN.

The SNN consists of two identical subnets with shared (identical) weights of connections. The main idea underlying the SNN is to train the subnets to compare a pair of feature vectors in terms of their semantic similarity or dissimilarity. One of the feature vectors from the pair is fed to the input of one subnet, and the second vector of the same structure is fed to the input of the second subnet. Semantically similar objects are defined by their belonging to the same diagnosis. Thus, the training process uses all possible pairs of available training examples from the dataset.

One of the peculiarities of the proposed approach is that two SNN are used to make a decision about a disease (see figure 1). The first SNN consists of two CNN such that direct images of segmented lung objects are fed to the input of the CNN. The second SNN uses histograms of the segmented objects as input data. Every pair of histograms is computed for every pair of images from the dataset. It is possible to collect a parallel dataset consisting of histograms to reduce the testing time.
Figure 1. An architecture of the CAD system.

In the testing phase, an image of an unknown new object is fed to the input of the first CNN. The second image for the input of the second CNN is taken from the dataset. If the neural network shows that these two images are semantically similar, then we make the diagnosis corresponding to the object from the dataset. Otherwise, the next image from the dataset is fed to the second input, and this comparison procedure is repeated. In parallel, the same testing process is carried out with histograms. As a result, we can compare the obtained results and to make a more confident diagnosis. Various strategies of decision making can be applied to the pair of the SNN.

An advantage of the proposed architecture is the partial implementation of the doctor’s decision making. Every SNN sequentially compares a new tissue with all available examples from the dataset. A doctor acts similarly by analyzing the CT image based on her/his experience, previously analyzed images.

5. Conclusion
The implementation of ideas underlying the proposed CAD system has been considered in the paper. The preliminary numerical experiments with the CAD system under condition of the incomplete dataset of atypical cases have illustrated its outperformance. At the same time, we have to point out that two main problems have to be solved in order to implement a CAD system of the highest accuracy. The first one is to significantly increase the amount of data in the dataset. The second problem is to improve a framework for solving the segmentation problem. These are directions for further research.

Acknowledgement
This work is supported by the Russian Science Foundation under grant 18-11-00078.
References

[1] Siegel R, Miller K and Jemal A 2018 *CA Cancer J. Clin.* 68 7–30
[2] Zhang J, Xia Y, Cui H and Zhang Y 2018 *Biomedical Signal Processing and Control* 43 138–147
[3] Zhang G, Jiang S, Yang Z, Gong L, Ma X, Zhou Z, Bao C and Liu Q 2018 *Computers in Biology and Medicine* 103 287–300
[4] Afshar P, Mohammadi A, Plataniotis K, Oikonomou A and Benali H 2018 From hand-crafted to deep learning-based cancer radiomics: Challenges and opportunities arXiv:1808.07954v1
[5] Cheplygina V, de Bruijne M and Pluim J 2018 Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis arXiv:1804.06353
[6] Litjens G, Kooi T, Bejnordi B, Setio A, Ciompi F, Ghafoorian M, van der Laak J, van Ginneken B and Sanchez C 2017 *Medical Image Analysis* 42 60–88
[7] Setio A et al 2017 Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge arXiv:1612.08012v4
[8] Brant W and Helms C 2012 *Fundamentals of Diagnostic Radiology* 4th ed (Philadelphia: Lippincott Williams & Wilkins)
[9] Fei-Fei L, Fergus R and Perona P 2006 *IEEE Trans. Pattern Analysis and Machine Intelligence* 28 594–611
[10] Koch G, Zemel R and Salakhutdinov R 2015 *Proceedings of the 32nd International Conference on Machine Learning* (Lille, France) 1–8
[11] Holzinger A, Biemann C, Pattichis C and Kell D 2017 What do we need to build explainable ai systems for the medical domain? arXiv:1712.09923
[12] Smith S and Jain A 1982 *Computer vision, graphics, and image processing* 20 259–271