Advances in Intelligent Systems Research, volume 174
Proceedings of the 8th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2020)

Identification of Pathological Formations in the Lungs Based on Machine Learning Methods

G. R. Shakhmametova
Computer Science and Robotics Department
Ufs State Aviation Technical University
Ufa, Russia
shakhgouzel@mail.ru

N.O. Vakkazov*
Computer Science and Robotics Department
Ufs State Aviation Technical University
Ufa, Russia
nikitavakkazov@gmail.com

R.Kh. Zulkarneev
Faculty of General Medicine
Bashkir State Medical University)
Ufa, Russia
zurustem@mail.ru

***Abstract*—The article discusses the use of deep neural networks for analysis and recognition of images of computed lung tomography. The recognition is carried out in two stages: segmentation of lung image on the CT slice and search of pathological entity. An algorithm for segmentation of lung images in images is proposed and a model for finding pathological formations based on machine learning methods is developed. To implement the stage of recognition of pathological formations, CNN convolutional neural network is used. The developed approach ensures that areas containing pathological formations are found on sections of pictures. Images from the public LIDC/IDRI database were used to test the model. The efficiency analysis showed on the test data the accuracy of the proposed model 0.82. The software is implemented in the programming language Python 3.6 and is cross-platform. For machine learning algorithms TensorFlow 1.14, Scikit-learn 0.22.1 packages were used.**

Keywords—machine learning, neural networks, CNN, computer tomography, lungs' pathological formations

I. INTRODUCTION

The introduction of IT technologies into healthcare now provides the ability to collect, store and process vast amounts of medical data that can be presented as numbers, text, images, videos, sound recordings. This means that the data are, in the vast majority of cases, heterogeneous and poorly structured. Machine learning techniques provide an effective set of tools for analyzing and processing virtually any type of data. The use of machine learning techniques to process medical data solves a number of problems related to diagnosis of diseases, monitoring the condition of the patient, predicting the condition of the patient, researching new drugs, and other problems. Data are obtained through various research methods, one of which is computed tomography.

Computed tomography (CT) is an imaging method based on the use of X-ray photons for imaging with the aid of digital reconstruction [1]. The detection of minor pathologies on computed tomography of a lung is an important problem of modern medicine, as the diagnosis of diseases in the early stages of development increases the effectiveness of further treatment and the chances of the patient's full recovery. But the early stages of diseases are often difficult to detect by visual analysis of tomograms. Analysis of tomograms and diagnosis of pathologies is carried out by a doctor radiologist, so the diagnosis is influenced by the human factor. In order to reduce the level of influence of the human factor, as well as the load on the diagnosticians, it is proposed to use artificial intelligence and machine learning methods for analysis and recognition of computed tomography images, in particular, deep neural networks, as one of the most effective tools for image recognition today [2]. Analysis and processing of CT snapshots is divided into several stages: preliminary processing, segmentation, recognition, and diagnostics [3]. Each stage has its own set of methods and algorithms.

This article describes the algorithm of detecting formations in lungs on CT sections, based on the mathematical model of deep learning. A public LIDC/IDRI database is used to train the model, containing computer lung tomograms with X-ray annotations. The second part presents existing CT image analysis and processing solutions. In the third part, the algorithm of recognizing pathological lung formations in CT images is considered. The fourth part presents the test results and analysis of the efficiency of the developed algorithm.

II. STATE OF ART

At present, numerous software products have been developed to visualize CT results, but many of them do not have the ability to detect pathologies in the lungs.

MeVisLab is the software for imaging, segmentation and image processing in medical research [4]. Among its advantages are the possibility to expand modular libraries of image processing and availability of a free version. Some disadvantages include absence of functions for pathology recognition in basic configuration.

AdvantageSim MD is the software for localization and simulation software [5]. Its advantages are possibilities of complex and multiphase modeling as well as automatic segmentation of bodies. Notable disadvantages are absence of functions for pathology recognition in basic configuration as well as absence of free version.

OncoQuant is the software for medical diagnostics [6]. Its advantages include solving problems of matching images

taken at different time intervals. Special automatic viewing protocols allow you to download similar series. It has the function of automatic matching of images obtained using different methods. Among its disadvantages are absence of functions for pathology recognition in the basic configuration, and the absence of a free version.

3DimViewer is an open source software product [7] supporting threshold-based tissue segmentation. It allows to model the surface of segmental tissue. This product is distributed for free. Its disadvantage is absence of functions for pathology recognition in basic configuration.

III. Proposed Solution

An efficient computed lung tomography (CT) picture consists of a number of images or sections. The number of sections depends on the characteristics of the tomograph. The evaluation of the tomogram is reduced to the task of segmentation followed by binary classification of the image areas.

In the CT (tomogram) image, the formations look like white spots (Fig.1).

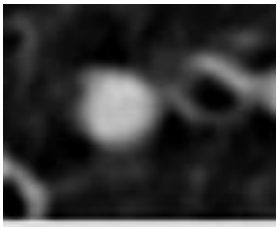

Fig. 1. Malignant formation (image on one of tomogram sections)

In order to detect pathological formations, it is necessary to process and recognize each section of the CT image. The process is carried out in two stages:

1) Segmentation of lung image on the CT slice.
2) Search (recognition) of pathological entity.

A. Segmentation of a lung image on a slice

Pathological formations are found only in the interior of the lungs. In order to increase the accuracy of the algorithm operation, it is necessary to cut off unnecessary areas of CT picture. Segmentation is carried out in several stages:

1) binarization of the image;
2) removal of borders;
3) selection of closed areas in an image and selection of the two largest areas;
4) elimination of noise in selected areas of the image;
5) overlaying the resulting binary mask on the original image.

The image is binarized at the specified threshold of 0.604. This threshold is is chosen for cutting off unnecessary structures on the image (bones, etc.). Initially, the threshold was taken from the authors' studies [8], but subsequently it showed good efficiency and thus remained unchanged.

Methods implemented in the skimag library [9] have been used to remove boundaries and highlight closed areas. The resulting regions correspond to an image of the inner surface of the lungs on the cut.

Once we have found the areas, we need to sort them by area and select the two largest ones. Thus, we select parts of the image pertaining to the inner surface of the lungs (areas in which it is necessary to search for pathological formations), as they occupy the most space on the image.

Next, you must adjust for any noise within the selected areas. Noise removal occurs in several steps:

1) Apply erosion operation with disk radius equal to 2.
2) Apply closing operation with disk radius equal to 9.
3) Find boundaries using Roberts cross-operator.
4) Fill all holes on binary mask.

Methods from skimage [9] have been used to perform this task.

The final step is to superimpose the mask obtained in the previous steps on the image that was before the segmentation. This is done by multiplying the binary mask by the image.

Fig.2 shows the result of operation of light image segmentation algorithm in CT image.

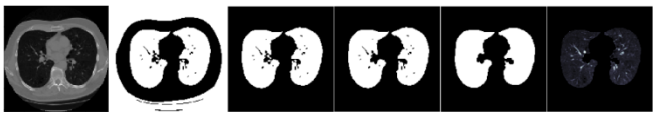

Fig. 2. Step-by-step segmentation on CT slice

B. Search (recognition) of pathological formations

Processing each section of a CT image is performed independently of others. The size of each slice is 512x512 pixels.

To search for formations, the slice is divided into 361 50x50 pixel areas (Fig. 3). Each region thus obtained is fed to the model input for binary classification.

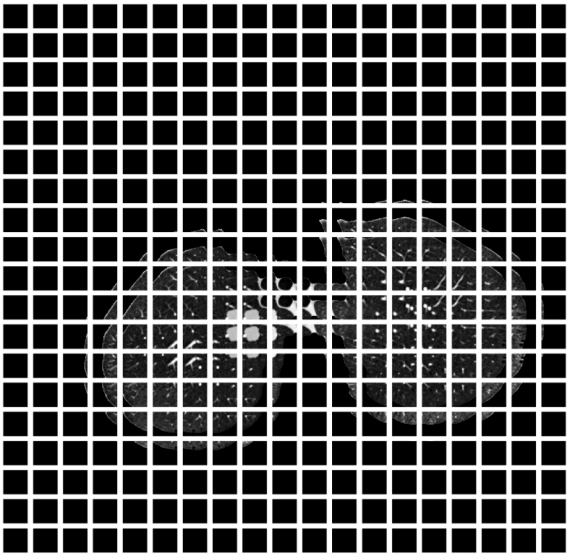

Fig. 3. Result of image division into areas

The model was a convolutional neural network (CNN) [10]. The choice is due to good image recognition results and acceptable complexity of calculations [11].

IV. Convolutional Neural Network Architecture

Convolutional Neural Network (CNN) is a special neural network architecture proposed by Jan Lecun, originally aimed at effective image recognition [11]. The model receives the 50x50x1 tensor at the input.

The model includes:

1) Convolution layer. Core 5x5, 32 filters. ReLU activation function. Output tensor 32x46x46x1.
2) Subsampling layer. Kernel 2x2. Output tensor 32x23x23x1.
3) Convolution layer. 5x5, 64 filter core. ReLU activation function. Output tensor 64x19x19x1.
4) Convolution layer. 3x3 core, 64 filters. ReLU activation function. Output tensor 64x17x17x1.
5) Subsampling layer. Kernel 2x2. Output tensor 64x8x8x1.
6) Full-knit layer. 512 neurons. Activation function ReLU.
7) Dropout layer. Coefficient of thinning 0.5.
8) Output layer. 2 neurons. Function of activation soft-max.

ReLU activation function:

$$f(x) = \begin{cases} 0, x < 0 \\ 1, x \geq 0 \end{cases} \quad (1)$$

Function of activation sof-max:

$$f(\vec{x}) = \frac{e^{x_i}}{\sum_{j=1}^{J} e^{x_j}} \quad (2)$$

Adam with a learning speed factor of 0.001 was chosen as the network optimizer. The Adam method converts the gradient as follows:

$$S_t = \alpha \cdot S_{t-1} + (1-\alpha) \cdot \nabla E_t^2;\ S_0 = 0 \quad (3)$$

$$D_t = \beta \cdot D_{t-1} + (1-\beta) \cdot \nabla E_t^2;\ D_0 = 0 \quad (4)$$

$$g_t = \frac{D_t}{1-\beta} \cdot \sqrt{\frac{1-\alpha}{S_t}} \quad (5)$$

$$\Delta W_t = \eta \cdot (g_t + \rho \cdot W_{t-1}) + \mu \cdot \Delta W_{t-1} \quad (6)$$

Categoric crossentropy is selected as loss function:

$$CE = -\log\left(\frac{e^{S_p}}{\sum_j^C e^{S_j}}\right) \quad (7)$$

At the output of the network we get the probability of belonging to class 1. Class 0 - no pathological formations were found on the image, class 1 - pathological formation on the image was found.

The training was conducted on images from the public LIDC/IDRI database [12]. The database contains data on 1018 CT results with annotations from X-ray specialists. Each CT result contains 100 to 600 image layers.

A sample of 9,797 50x50 images was formed to train the model. The sample was divided into learning (6878 images), validation (1297 images), and test (1622 images). Each image has a class label of 0 or 1.

Fig.4 is a graph showing the accuracy metric of algorithm operation on validation data at each training step.

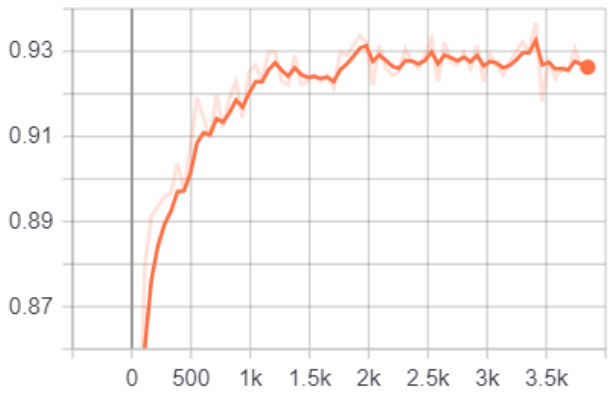

Fig. 4. Accuracy

Fig. 5 is a graph showing the value of validation data loss at each training step.

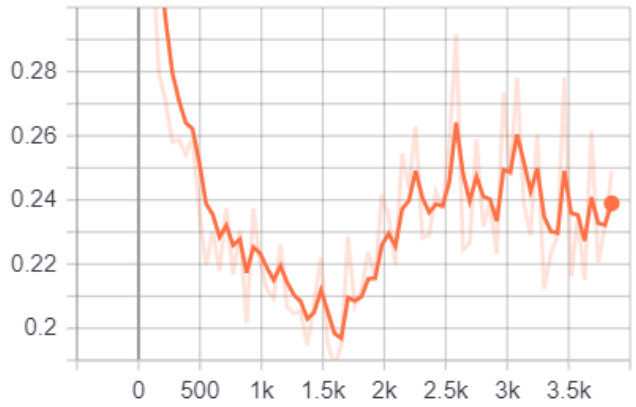

Fig. 5. Loss

The best model turned out at 1650 step of training. In Fig.5 this step corresponds to minimum loss value.

V. Results

Images from the public LIDC/IDRI database were used to test the model. Testing was performed on a sample of 1,622 50x50 pixel images. Of these, 282 images were Class 1 (there is education) and 1340 images were Class 0 (there are no formations). Fig.6 shows the error matrix.

Analysis of the error matrix shows that 1,521 images are classified correctly (of which 1,304 have no formations and 217 have formations) and 101 images are classified incorrectly (of which 36 images the model classified as

containing formations and 65 images marked as containing no formations)

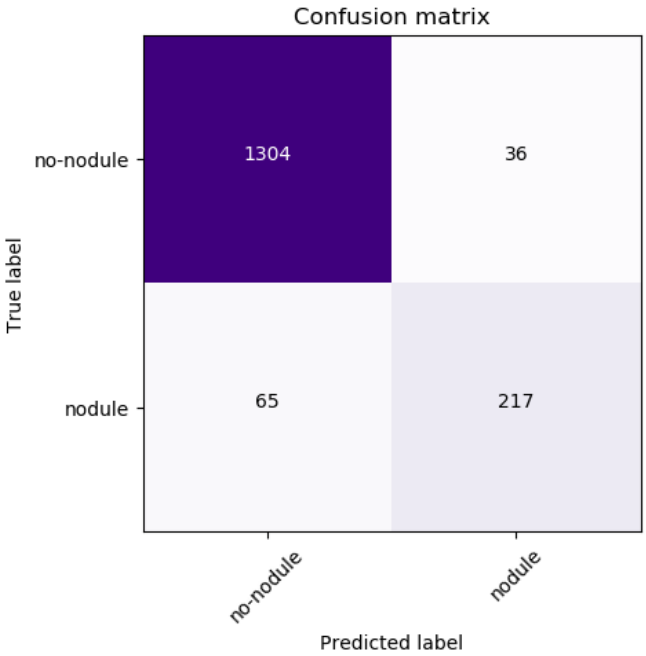

Fig. 6. Errors matrix

Calculation of metrics:

$$P = \frac{217}{217 + 36} = 0.86 \quad (8)$$

$$R = \frac{217}{217 + 65} = 0.77 \quad (9)$$

$$F = 2 * \frac{0.86 * 0.77}{0.86 + 0.77} = 0.82 \quad (10)$$

Since the sample is unbalanced, we use F metric. Thus, the accuracy of the model is 0.82 on the test data.

$$Type\ I\ errors = \frac{36}{1622} = 0.022 \quad (11)$$

$$Type\ II\ errors = \frac{65}{1622} = 0.04 \quad (12)$$

The result of algorithm operation is shown in Fig.7. It shows segmented lungs and a frame with the specified probability of being in it pathological formation (0.989).

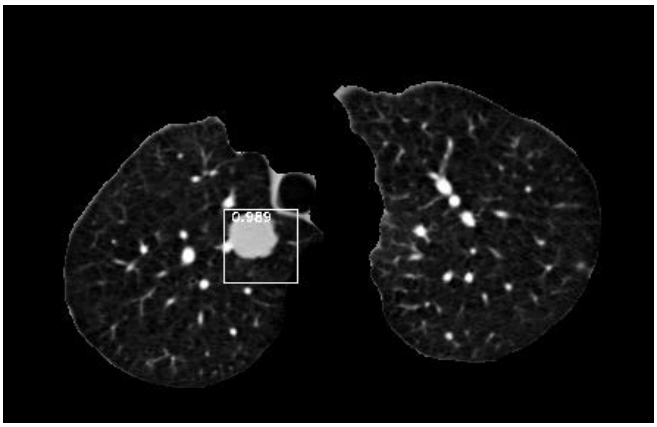

Fig. 7. Result of operation for the algorithm of formation detection in lungs

The developed software identifies 50x50 pixel areas on each slice where the probability of having formations above the 0.9 threshold is set and displays this value in the upper left corner of the frame.

The software is implemented in the Python 3.6 programming language. The interface is designed with the GUI PyQt 5 module. For machine learning algorithms, TensorFlow 1.14, Scikit-learn 0.22.1 packages were used. Scikit-image 0.17.2, Pillow 7.1, OpenCV 3.4.2 packages were used for image processing algorithms.

Figure 8 shows a main screen picture of developed software.

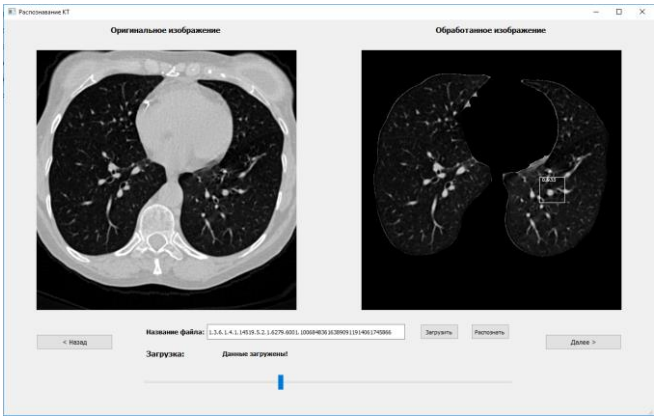

Fig. 8. Software main screen

We test efficiency on one of the samples of CT shot (312 layers). Table 1 provides information on errors in recognition of pathological formations

TABLE I. RECOGNITION ERRORS

		Predict label	
		False	***True***
True label	False	34008	21
	True	10	28

Next we will calculate the errors of the 1st and 2nd type:

$$Type\ I\ errors = \frac{21}{34008} = 0.0006 \quad (13)$$

$$Type\ II\ errors = \frac{10}{34008} = 0.0003 \quad (14)$$

Thus, the proposed model found 73% of the areas in which pathological formations were located. 42% of areas marked as containing pathologies turned out to be false. From this it can be concluded that to date the model is unable to completely replace the specialist radiologist, but can provide significant assistance in supporting decision-making when detecting areas in the lungs in which pathologies can be found with a high probability, including malignant ones in the early stages.

VI. Conclusion

The result of the study is the development of a model for detecting pathological formations in the lungs, including the early stages, from images of computed lung tomography based on machine learning techniques. The implementation of the model involves 2 stages: segmentation of the lung image on the cut to increase the accuracy of recognition and search (recognition) of pathological formations. To implement the stage of recognition of pathological formations, CNN convolutional neural network is used, architecture is developed and network training is carried out. The software is implemented in the programming language Python 3.6. The interface is developed with the help of the module for creation of GUI PyQt 5. For machine learning algorithms, TensorFlow 1.14, Scikit-learn 0.22.1 packages were used. For image processing algorithms, Scikit-image 0.17.2, Pillow 7.1, OpenCV 3.4.2 packages were used. The test results showed a fairly high recognition accuracy of 82%. The software developed is cross-platform and can provide significant assistance to the diagnosticians in supporting their decision-making when recognizing pathological formations in the lungs.

Further research in this area is planned in the field of improving the accuracy of recognition for pathological formations and recognition of other lung diseases from CT images.

Acknowledgment

The reported study was funded by RFBR according to the research projects № 19-07-00780, 19-07-00709.

References

[1] Computed tomography [Electronic Resource], URL: https://www.who.int/diagnostic_imaging/imaging_modalities/dim_comptomography/ru/ (date of the request: 10.05.20).

[2] Nikolenko S., Kadurin A., Archangelskaya E. Deep learning. Immersion in the World of Neural Networks, SPb.: Piter, 2020, 480 p.

[3] Doronicheva A.V., Savin S.Z. (2014). Methods of recognition of medical images for computer automated diagnostics tasks. Modern problems of science and education, 2014, No. 4. URL: http://www.science-education.ru/ru/article/view?id=14414 (date of the request: 10.05.20).

[4] Medical Image Processing and Visualization [Electronic Resource], URL: https://www.mevislab.de/ (date of the request: 10.05.20).

[5] Advanced localization and simulation software to improve and optimize radiotherapy planning [Electronic Resource], URL: https://www.gehealthcare.ru/products/advanced-visualization/all-applications/advantagesim-md (date of the request: 10.05.20).

[6] A set of robust tools for conducting standard cancer diagnosis procedures and subsequent control studies [Electronic Resource], URL: https://www.gehealthcare.ru/products/advanced-visualization/all-applications/oncoquant (date of the request: 10.05.20).

[7] Official site of 3Dim Laboratory [Electronic Resource], URL: https://www.3dim-laboratory.cz/en/software/3dimviewer (date of the request: 10.05.20).

[8] Armato III, Samuel & Macmahon, Heber. (2003). Automated lung segmentation and computer-aided diagnosis for thoracic CT scans. International Congress Series. 1256. 977-982. DOI: 10.1016/S0531-5131(03)00388-1 (date of the request: 10.05.20).

[9] Image Processing for Python [Electronic Resource], URL: https://scikit-image.org/docs/dev/api/skimage.html (date of the request: 10.05.20).

[10] Farhana Sultana, Abu Sufian, Paramartha Dutta. (2018) Advancements in Image Classification using Convolutional Neural Network. Fourth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN'2018) DOI: 10.1109/ICRCICN.2018.8718718 (date of the request: 10.05.20).

[11] Jianxin Wu. Convolutional neural networks [Electronic Resource], URL: https://cs.nju.edu.cn/wujx/teaching/15_CNN.pdf (date of the request: 10.05.20).

[12] LIDC-IDRI [Electronic Resource], URL: https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI (date of the request: 10.05.20).

.