AIzimov: the Platform for Intellectual Diagnostics of Lung Cancer

Alexey Lukashin, Alexander Ilyashenko, Lev Utkin, Vladimir Muliukha
Peter the Great St. Petersburg polytechnic university
St. Petersburg, Russian Federation
alexey.lukashin@spbstu.ru, ilyashenko.alex@gmail.com, lev.utkin@mail.ru, vladimir@mail.neva.ru

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

***Abstract*—The paper describes the practical approach of building a platform for collaboration between doctors, computer and data scientists. This study is related to a research project of creating intellectual methods of lung cancer detection which includes building new data sets, development of new methods and creating software for data collection and processing. The paper overviews solution architecture, algorithms, and software tools for processing computer tomography data, providing remote access to the user groups, data visualization, and processing on supercomputer systems.**

I. Introduction

For the last years intellectual diagnostics which applies machine learning methods for helping medical personnel to take a decision become extremely popular. Studies about development of new methods and algorithms cover a lot of different topics, but medicine is one of the most popular. There are a lot of research projects related to cancer detection using machine learning methods [1], [2]. This paper is about one of such projects [3] dedicated to lung cancer detection based on intellectual processing of computer tomography (CT) scans. This project is running in Peter the Great St. Petersburg polytechnic university with a strong help of doctors from St. Petersburg oncology center. The project team is becoming bigger and there are new organizations like Rostov oncology center are joining to this project.

The motivation of this work is following. If on the project is working only data science team and running experiments on public data sets (e.g., for lung cancer there are several public data sets available like LUNA16 or LIDC) then platform and tools are not required. But if project becomes multidisciplinary and there are a lot of distributed team members which work on different topics like data science, software development, data collection, interpretation then the ecosystem of connected tools is strongly required. For supporting this project authors have done a lot of work not directly related to the intellectual methods development. This work includes following topics:

1) Organizing data collection and markup and providing doctors useful tools for making data markup easier;

2) Organizing remote work of different geographically distributed teams on the same data sets;

3) Implementing of data processing on supercomputer systems for training and data inference;

4) Providing online visualization of CT scans and markup tools as a web application;

5) Ensuring access control to the data, CT scans anonymization, and making own data set publicly available;

6) Storing results of training (models) with metadata on which data set model was trained, model precision on the test set, and other parameters.

The developed tools and solutions are integrated into one platform which is called AIzimov. This platform allows to organize collaboration between doctors and data scientists and provide convenient way of data gathering and collection. This paper describes challenges which faced project team and technical details of developed algorithms and solutions

The paper is organized into seven sections. First section is introduction. Second section overviews project details, short description of intellectual diagnostics approach and references to the related work. Third section describes general architecture of the AIzimov platform and its main components. Forth section overviews solutions for data gathering from geographically distributed locations and providing doctors tools for processing this data. Fifth section is describing challenges for providing remote viewer of CT scans for doing data markup and showing results of intellectual diagnostics. Sixth section contains algorithms and technical solutions for processing data on supercomputer systems. The last, conclusion section overviews results of this work and states topics for the further research and development.

II. Overview of Lung Cancer Detection Project

Nowadays, medicine and computer technologies are working together. Computers are used in surgery for operations, in diagnostic procedures and analyzing the results. However, their usage is far from possible everywhere or difficult. Many procedures are still carried out manually by doctors. With the growing interest to machine learning and the rapid development of new computer hardware in the last decade, it

has become possible to build assistants for doctors based on machine learning models to speed up and simplify diagnostic process. This is not a complete replacement of the doctor, but only help in analyzing any results. In modern medicine, almost every clinic is equipped by tomographs and can serve hundreds of patients a day. However, such highly intensive flow of patients requires a large number of diagnosticians who will carefully review these tomographies. In particular, a doctor spends from tens of minutes to several hours analyzing a single tomography. So, appearance of an intellectual assistant that will identify typical cases for which he was trained will allow to reduce processing time of single tomography and increase throughput.

The process of automated intellectual analysis of tomography in case of lung cancer can be separated into following steps: preparation, segmentation and classification [7]. The system which implement diagnostic processes is called CAD (computer-aided diagnostics). This system is split into two subsystems: CADe subsystem which is responsible for preparation and segmentation and CADx system which is performing classification of nodules.

Preparation step is step when tomography slices are getting filtered by specific rules and splitted into multiple images for different segmentation rules. Then each image is passing procedure of segmentation to find each feature which looks suspicious. Then each suspicious area passing throw multiple classificators, built with using Siamese neural networks and weighted random forest models [5], [6], to classify them as malignant or benign.

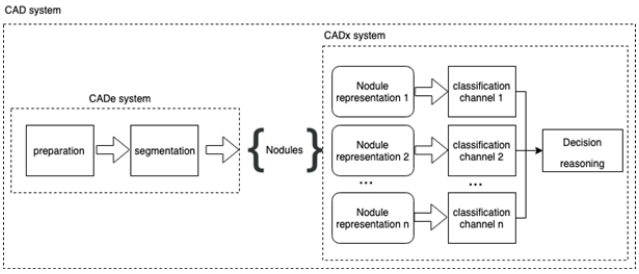

Fig. 1. CAD system with multichannel CADx subsystem

The proposed approach (Fig. 1) means that any CT scan is processed using different classification channels for improving precision and accuracy of classification. Different channels serve different purposes. For example, channel which applies nodule classification using Siamese or Triplet neural network is able to detect atypical cases, and channel which applies classification using weighted random forest models is able to detect typical cases with good precision.

III. AIZIMOV PLATFORM GENERAL ARCHITECTURE

The platform is called AIZimov, sounds like to the name of famous writer Isaac Asimov, who was inspiring project authors for creating new intellectual methods in tools for medicine and robotics (cyberphysical) applications. The platform includes following systems:

1) Main system which is deployed to the infrastructure of the St. Petersburg polytechnic university and is responsible for keeping and processing data, providing services and applications to the end users;

2) Partner organization system. This system is installed in the hospitals or other partner organizations to provide ability to work with local copy of data and upload new data sets to the main system.

The architecture of the main system is shown on the diagram below (Fig. 2):

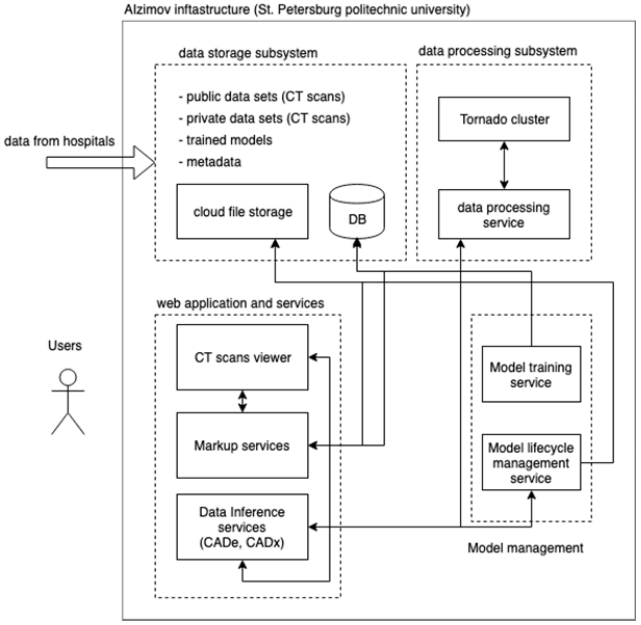

Fig. 2. Architecture of the main AIzimov system

The main system consists of following subsystems:

A. Data storage subsystem

Data storage subsystem consists of file storage with cloud architecture and relational database. Cloud storage is used to keep large files like CT scans and files with CT scans markup in organized folder structure. The relational database is used for keeping metadata information, user accounts, links between data files, etc. For providing easy data access for a file storage owncloud service is used.

B. Data processing subsystem

Data processing subsystem includes service which is connected with Tornado cluster. Tornado cluster is the one of the supercomputers located in St. Petersburg supercomputer center. The integration between AIZimov platform and computation cluster is presented in Section VI.

C. Web application and services

This subsystem is containing services for working with platform users. It includes data inference services which take CT scan as an input and process it using data processing

services. I also includes web application (viewer for CT scans) and a special markup service for annotating CT scans (markup of tumor) and keeping an annotation metadata in the data storage subsystem. The specifics of these services are overviewed in Section V.

D. Model management subsystem

This subsystem is intended for use for training models and working with datasets. Services of this subsystem allow to define a data set by selecting data from data storage subsystem and run a model training process. A trained model and its metadata (link to dataset, quality parameters) are stored in data storage subsystem. An important functionality of this subsystem is versioning of models. For this DVC tool (data version control) is used [4].

Partner organization system is setup close to the data sources and experts. This is mainly hospitals. The first setup was done in St. Petersburg oncology center. The architecture of this system is shown on Fig. 3.

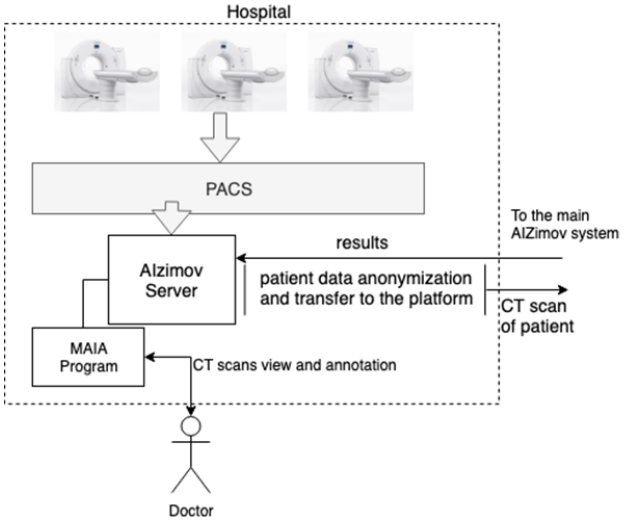

Fig. 3 Architecture of the AIzimov system which is deployed to partner organisation

On the diagram there are following components:

A. Medical equipment

AIzimov is integrated with PACS systems (Picture Archiving and Communication System). When a new CT scan is coming to PACS from tomography medical equipment it is automatically fetched to local AIzimov server.

B. AIzimov systems

AIZimov server is responsible for communication with PACS systems (getting new CT scans), data anonymization and annotation, and working with main AIzimov platform. To provide a convenient way of data annotation there is another software tool called MAIA (medical artificial intelligence assistant). This tool allows doctor to open CT scan locally, perform annotation and send it to the global data storage through AIzimov server. MAIA program repeats well known for medical personnel applications for viewing CT scans in DICOM format. The interface of the program is shown on the Fig.4:

Fig. 4. MAIA annotation tool multi planar view interface

IV. DATA GATHERING APPROACH FROM GEOGRAPHICALLY DISTRIBUTED MEDICAL CENTERS

First medical organization which was connected to the AIzimov platform is Clinical Research Center of Specialized Types of Medical Care (oncological) located in St. Petersburg. Because of the same geographical location it was easy to setup AIzimov components on the side of oncological center and tune data flows. CT scans are quite large in terms of data volume. An average size of a CT scan is about 500 megabytes and the database of couple hundreds of images might consume tens of gigabytes. The next medical center which is being connected to the AIzimov infrastructure is Rostov-on-Don oncological center. This organization is about 1800 kilometers away from St. Petersburg and setting up collaboration tools became not so easy.

For solving the integration issue following classification of collaboration types is proposed:

1) Full integration for data providers. This type of integration assumes full setup of AIzimov services on the partner's side. It includes AIzimov server, synchronization with cloud storage, MAIA software. It also includes full access to the web application. This setup is mainly required for the organization which is generating data (annotated CT scans) for its further delivery to the AIzimov data sets. Usually doctors work with data on the local infrastructure and sending ready data to the platform.

2) Remote work with data for annotation. This type of integration includes web application access and cloud storage access for working with local copies of data. This setup required for organizations which help to annotate data which is already located in the platform or have small data sets to work

on. This type of setup is useful when organization is located in another region or country.

3) Remote work with CT scans: running cancer detection and testing. This type of integration is intended for use when only remote data testing is required. It allows to run processes of intellectual diagnostics on CT images and viewing results. It requires only setting up new accounts and access restrictions in the web application.

Collaboration models listed above two types of working with data: locally and remotely. The goal of the platform is to provide ability to work with the same data using any preferable way. Another goal of the platform is security. Access to the data sets should be restricted and from the other side there is should be a way to quickly share data between different parties to simplify collaboration. Base on this points following structure of the cloud storage is proposed:

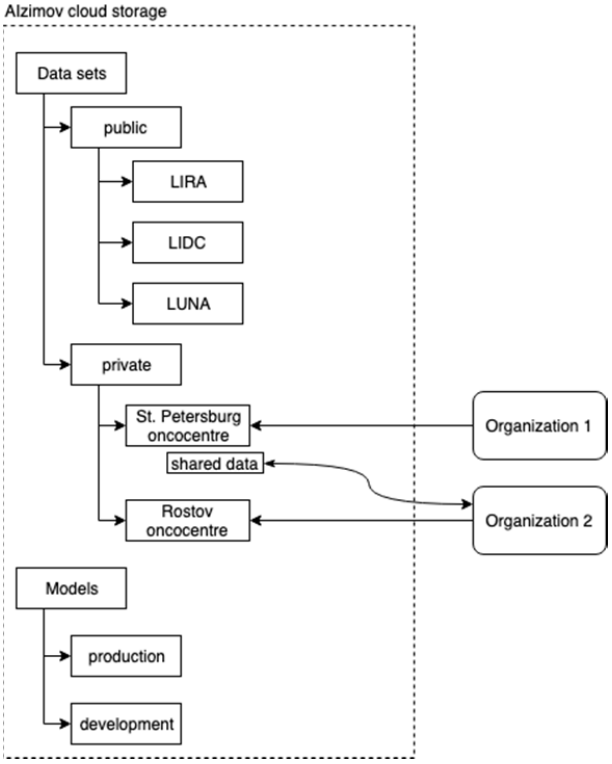

Fig. 5. The AIzimov cloud storage structure

On the Fig. 5 presented a structure of a cloud storage. Each item or sub-item has an owner. An owner can share access to the other platform users. It allows organizations to work independently from each other on own data sets. Such approach also allows to share subfolder with other party and work on the same data which is useful, for instance, for cross validation of annotated CT scans. There is also a separate data folder for models. This folder contains trained models which are used for the data inference. Each model has metadata: on which data set it was trained, on which data set it was tested, and model quality parameters. While working with AIzimov services which provide processes of intellectual diagnostics it possible to specify which model to use for cancer detection.

V. CT SCANS VISUALIZATION AND MARKUP IN WEB

Development of a web platform for processing computer tomography can solve a number of problems associated with the organization of the data gathering process. In the case of the development of other technical solutions necessary to overcome a lot of problems:

- the need to install applications in the infrastructures of hospitals and medical centers;
- organization of local data storage for computer tomographies;
- organization of testing and technical support;
- technical documentation development for system administrators and hospitals and medical centers staff training;
- organizing high-speed communication channels for sending data for analysis and for obtaining results.

All these limitations make it possible to opt for the web-based applications with secure communication channels for data transfer and anonymized storing. The most serious limitation in the development of such a solution is the size of the transferred data. Since computed tomography itself is a set of files with images, they are quite large. Each image is a slice obtained by scanning the patient and stores information about the received x-ray image.

Fig. 6. Axial plane for tomography viewer

For working with computed tomography in most cases, only the initial images obtained with a tomography medical equipment are not enough (Fig. 6). To process images, the doctor needs to be able to view tomography from three planes, which is called Multi Planar Reconstruction (MPR) and presented on Fig. 7. To do this, when loading a tomography,

you must completely load it into the computer's memory and build slices not only in axial plane but in two other planes – coronal and sagittal.

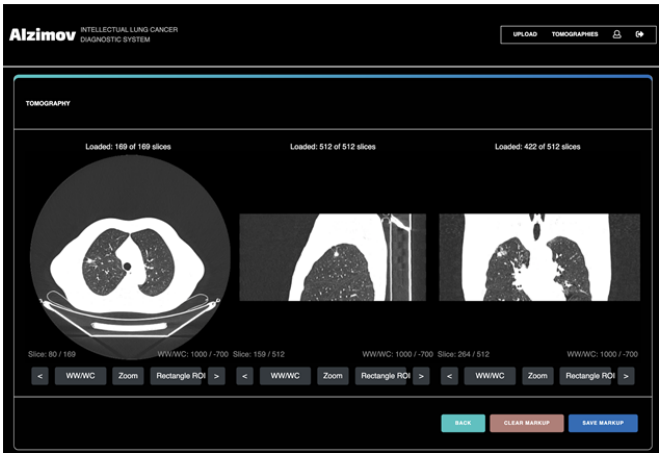

Fig. 7. Multi-planar reconstruction of tomography image

In this mode, doctor can view tomography not only in axial plane, and it becomes easier to determine the volume of growths and to evaluate the prevalence of process.

For working with tomographies we developed a client-server application for processing, displaying and marking of computer tomography. For the server part, Python language and the pydicom library were used, which allowed reading information from dicom files and receiving information from all slices. Then, after reading data from all axial slices into memory and using this library, all slices for sagittal and coronal planes can be built. This process is executed only once when user is loading his tomography to the server. And for each tomography in system we keep three subfolders with all slices for three planes and building them only once.

Client-side application allows to show CT, mark up CT and inform patients about found growths. For client side application development were used language JavaScript, framework React and library for working with slices Cornerstone.js [8]. When displaying slices, data is downloaded frame by frame, and it allows the doctor to start working with slices immediately from the beginning of work with tomography with updating data as tomography is loaded in all three planes.

Marking tomography in MPR mode, users can change a width of window by density for filtering tomography image, zoom in or out image and mark image using rectangular areas. Prompting the user for allocation of the rectangle on one of the planes stripes appear in the other views, showing level where rectangle was selected (Fig. 8). When a rectangle is selected on the second plane, a selected area appears in the third image showing the volume of the marked growth (Fig. 9).

Then user can save marked area with parameters of type of new growth, diagnosis, comments, method for verification and start searching for other growths. A list of all found growths is displayed at the bottom of the page in a table with the ability to select and search by type of growth.

Also, using React with JavaScript and adaptive layout we adapted this client application to be used on mobile devices, that doctor can use it from tablets and mobile phones.

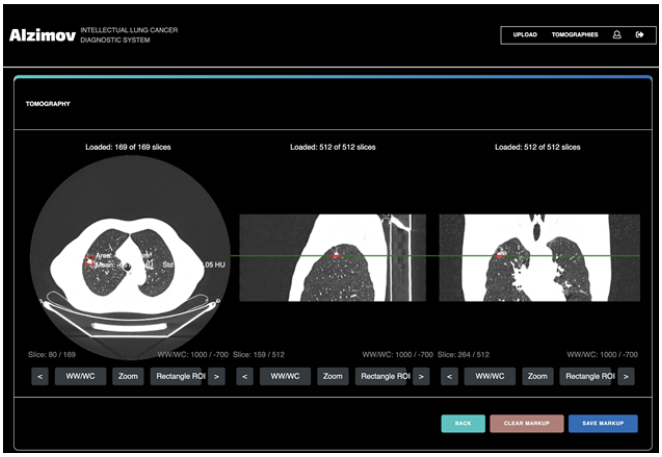

Fig. 8. Selected first rectangle area on axial plane and prompting on other two planes

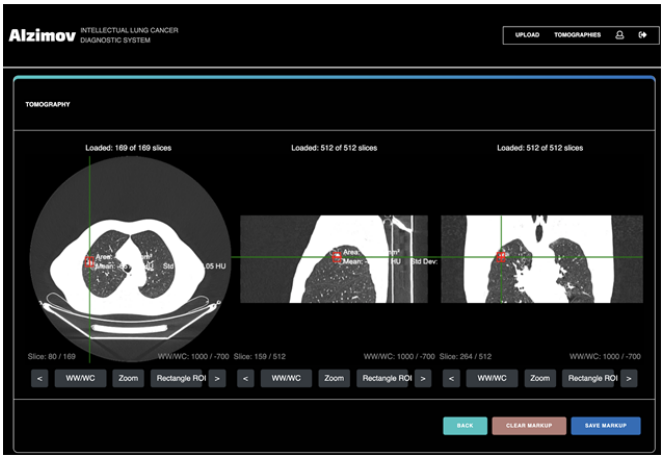

Fig. 9. Selected rectangle areas on two planes of tomography and showing volume of selected area

During the application development process, we consulted with specialists from the oncology center of St. Petersburg and discussed features of specialists working with tomography images.

VI. DATA PROCESSING ON SUPERCOMPUTER SYSTEMS

Computer tomographies are large amounts of data and their processing require high performance processing resources. Potentially this service can be used by a large number of users from various medical centers and incoming flow of tomography will be large and highly intensive. So large computational resources will be required to process them. For their fast processing, high-performance multi-core processors and graphics accelerators have to be used. But organization of such systems in each hospital and medical center may be too costly and require the need for qualified specialists to service such a system. The widespread practice of transferring such services to cloud systems and data centers can solve this problem. However, for processing tomography it is also

necessary to have high-performance systems for processing them. For this purpose, the resources of supercomputer centers can be used. In this platform, we used the resources of the supercomputer "Polytechnic" Center for processing and intelligent analysis of images.

The Polytechnic Supercomputer Center (SCC) located in Peter the Great St. Petersburg polytechnic university is one of the largest supercomputing centers in Russia, currently it is on fifth position in rating top50.supercomputers.ru. It consists of five different systems intended for use for numerical simulations, machine learning, big data analytics, etc. In this project for calculations, a Tornado cluster was selected that provides x86_64-based computational nodes with Intel Xeon E5 2697 v3 processors (28 cores, 56 hardware streams) and 64 GB of RAM and Nvidia Tesla GPUs. In total, there are 612 calculators of this type in the cluster and a productivity of more than 1 petaflops. Supercomputer systems are connected to a single network storage accessed through the SLURM (Simple Linux Universal Resource Manager) task scheduler. Using this open source scheduler allows you to develop software for task management and distribute them to computation nodes in accordance with the specified characteristics.

To run tasks on a supercomputer, you need to upload tomography data to run the task on one or multiple nodes of the supercomputer. To do this, over a secure connection via ssh, the archive with images is getting uploaded to the terminal server and then task with uploaded data is placing in the queue for execution. To organize the work of the queue, the SLURM task scheduler is used. It allows to control resources and to track statuses of tasks. However, it has a number of disadvantages described in [9]. To solve these problems we used the algorithm from [9], [10], which was developed for better distribution of supercomputer resources. The server part of the application periodically polls the task scheduler for tasks and determines a moment of time when results are ready and can be downloaded for further use (Fig. 10).

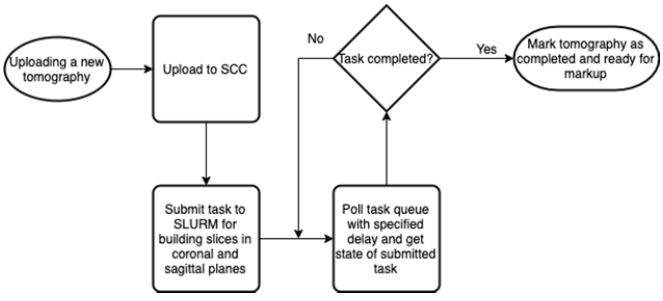

Fig. 10. Algorithm of initial tomography processing on SCC “Polytechnic”

Also, supercomputer is used in our system for training Siamese neural networks and weighted random forest models [5], [6] for intellectual classification of segmented slices and marking up tomographies. For training we always use results which were obtained from doctors working with our system when they markup tomography manually. We’re saving all these results for more accurate training neural network and retrain our system to make it better and increase an amount of true positive results. Algorithm of working with machine learning models similar to algorithm on Fig. 10 except submitting task that changes to submitting machine learning script.

Using a supercomputer allows to perform many tasks in parallel and process a large number of tomographies simultaneously using multi-core processors and graphic accelerators. Data processing is organized anonymously, thus the privacy of patients and their data is not violated and data transfer is done using encryption and secure data transfer protocols.

VII. Conclusion

In this paper an approach for building a platform for collaboration between data providers (medical centers) and intellectual system for diagnostics of lung cancer is presented. Platform organization allows to work on the datasets and share results of work in quick and easy way. Currently platform offers two ways of working with data: local and remote. There is no decision which way is preferable yet, but it’s on discussion with specialists from different medical centers.

The proposed architecture includes main system structure, services which have to be installed in the hospitals and other partner organizations, data structure in the cloud storage system, and details of services which provide ability to online work with CT scans through web application and integration with the supercomputer center Polytechnic for processing large amount of data.

The main challenge in a platform part of this project is size of CT images and data sets. It makes difficult providing fast and convenient way of work if partner organization has slow connection.

The proposed architecture is implemented in working software tools but still changing because of dynamic nature of this project. The further plans are to expand functionality of tools for data scientists and integrate more organizations and partners to the platform

Acknowledgment

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

The results of the study 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] J. Zhang, Y. Xia, H. Cui, and Y. Zhang, “Pulmonary nodule detection in medical images: A survey,” *Biomedical Signal Processing and Control*, vol. 43, pp. 138–147, 2018.

[2] G. Zhang, S. Jiang, Z. Yang, L. Gong, X. Ma, Z. Zhou, C. Bao, and Q. Liu, “Automatic nodule detection for lung cancer in ct images: A review,” *Computers in Biology and Medicine*, vol. 103, pp. 287–300, 2018.

[3] 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

[4] Open-source Version Control System for Machine Learning Projects, Web: https://dvc.org/
[5] L.V. Utkin, A.V. Konstantinov, V.S. Chukanov, M.V. Kots, M.A. Ryabinin, A.A. Meldo, "A weighted random survival forest", *Knowledge-Based Systems*, vol.177, Aug. 2019, pp. 136-144
[6] A.A. Meldo, L.V. Utkin, "A new approach to differential lung diagnosis with CT scans based on the Siamese neural network", *J. Phys.: Conf.*, Ser. 1236, 2019, n. 012058
[7] A.A. Meldo, L.V. Utkin, "Radiomics as a basis for transformation of radiologists skills and partnership", *J. Phys.: Conf.*, 2019, Ser. 1236 n. 012063
[8] Cornerstone.JS library official website, Web: https://cornerstonejs.org/
[9] Algorithms for planning Resource-Intensive computing tasks in a hybrid supercomputer environment for simulating the characteristics of a quantum rotation sensor and performing engineering calculations, Ilyashenko, A.S., Lukashin, A.A., Zaborovsky, V.S., Lukashin, A.A. // *Automatic Control and Computer Sciences*, 2017, 51 (6), pp. 426-434
[10] Methodology of Effective Task Planning and Algorithm for Multivariate Computation of the Characteristics of a Quantum Rotation Sensor on a Hybrid Supercomputer Cluster, Ilyashenko, A.S., Voskoboynikov, S.P., Ustinov, S.M., Lukashin, A.A. // *Automatic Control and Computer Sciences*, 2018, vol. 52, n. 6, pp. 496-504