DEVELOPMENT OF NEURAL NETWORK MODELS FOR OBTAINING INFORMATION ABOUT NODULAR NEOPLASMS OF THE THYROID GLAND BASED ON ULTRASOUND IMAGES

ILYA LOZHKIN[1], KSENIA TSYGULEVA[2], KONSTANTIN ZAYTSEV[3], MAXIM DUNAEV[4], SVETLANA ZAKHAROVA[5], EKATERINA TROSHINA[6], ALEKSANDER GARMASH[7]

[1,2,3,4,7]National Research Nuclear University MEPhI (Moscow Engineering Physics Institute), 31 Kashirskoe Avenue, Moscow, 115409, Russia

[5,6]National Medical Research Center for Endocrinology, Ministry of Health of Russia, 11 Dmitry Ulyanov str., Moscow, 117292, Russia

E-mail: kszajtsev@gmail.com

ABSTRACT

Analysis in ultrasound examinations of the thyroid gland often requires a large amount of training data, therefore, one of the important problems of using deep architectures in medicine is the expansion of training datasets to more significant volumes, since in most clinics the sets of such data are not large. The purpose of this study is to develop a set of neural network models for solving the problems of analyzing ultrasound images of the thyroid gland to identify nodular neoplasms, as well as to study various approaches to increasing the amount of data for the formation of training samples. According to the results of the conducted experiments, it was found that an increase in the number of the same type of images in film loops did not affect the operation of deep architectures, and therefore it was meaningless. Various approaches to the augmentation of medical image sets were investigated, and it was observed that the complication of the augmentation process of images containing enlarged areas with useful information negatively affected the quality indicators of segmentation models trained on such data. Based on the conducted research, the authors propose neural network models to solve the problems of semantic segmentation and classification for use in the field of ultrasound examinations of the thyroid gland. The results obtained allow for advancing the use of artificial intelligence methods for personalized medicine for thyroid diseases.

Keywords: *Machine Learning, Ultrasound Imaging, Detection, Segmentation, Classification.*

1. INTRODUCTION

In recent years, deep neural network architectures have provided unique opportunities and notable breakthroughs in solving problems in various fields [1, 2]. In clinical medicine, it has been shown that computer diagnostic systems based on deep learning provide competitive and sometimes even superior diagnostic accuracy and efficiency compared to experienced clinicians [3-6]. Although the application of deep learning techniques to medical images is usually used to improve diagnostic efficiency, there is an ever-growing demand for more advanced deep neural networks to solve new more complex scenarios.

Currently, many existing studies emphasize the competitiveness of deep convolutional neural networks (CNN) when they are used to diagnose various diseases. These models give decent results in the tasks of classification, detection, and segmentation to reveal a specific disease; however, in many cases, they are still inferior to the human-level diagnosis. This is because experienced clinicians usually use additional domain knowledge to make a diagnostic conclusion, rather than relying solely on medical images. Therefore, deep neural networks analyzing images can only make assisting conclusions in the field of ultrasound diagnostics. However, they can successfully detect interesting areas in the image or segment nodes and other neoplasms.

The use of neural networks in the field of thyroid ultrasound is critically important for accurate diagnosis and effective treatment of patients. Neural networks are the most effective tool for automatic processing and analysis of medical images. Their use allows for automatically determining the shape, size, structure, density, and other important characteristics of the thyroid gland. Neural networks can also track changes in the thyroid gland at different stages of its development, which is an important factor in predicting possible diseases.

Automation of the process of thyroid nodule ultrasound diagnostics is a complex task. This conclusion is based on the method of detecting nodular formations by doctors: to make a diagnosis, they must first detect (localize) the node, then isolate and analyze (segment) its outlines, and finally classify and describe it [7].

The key stage in the diagnostic process is to identify the outlines of the nodular formation and determine its type. Therefore, the tasks of segmentation and classification of ultrasound images must be fully solved.

The implementation of algorithms must meet the following requirements:

1. Independence of the result from the input data. The model must be trained in such a way that the format, location, framing, and resolution of the input data do not affect both additional training of the model and predictions during testing and use.

2. Independence of the result from the capabilities of diagnostic devices. Since different medical institutions use different equipment, segmentation, and classification models should be mobile. The algorithm must be trained in such a way that it is ready for any changes in the data. To meet this requirement, it is proposed to expand the training dataset for the algorithm with augmented data.

3. A comprehensive solution to the problem. The task of segmentation and classification of such complex images as ultrasound images is not trivial. Since the result of the work can be influenced not only by the type of image (longitudinal or transverse) during the study, but also by the type of nodular formation, neck width, etc., it is necessary to provide an approach to the implementation of segmentation to minimize the influence of the described circumstances on the diagnostic process.

4. Minimizing the requirements for manual data markup. Since compliance with the requirements described above is influenced by many factors, this paper describes the study of the influence of training data characteristics on the results of detection, classification, and segmentation algorithms, as well as the creation and training of universal medical image analysis solutions.

Thus, the use of neural networks in the field of ultrasound examinations of the thyroid gland is necessary to improve the accuracy of diagnosis and prediction of diseases. They are universal for image processing approaches and can be integrated with data from various sources. Analysis in ultrasound examinations of the thyroid gland often requires a large amount of training data. However, in this case, the data usually have a small volume and are quite expensive to obtain. Therefore, one of the important problems of using deep architectures in medicine is the expansion of training datasets to more significant volumes, since in most clinics the sets of such data are not large.

The purpose of this study is to develop a set of neural network models for solving the problems of analyzing ultrasound images of the thyroid gland to identify nodular neoplasms, as well as to study various approaches to increasing the amount of data for the formation of training samples.

2. MATERIALS AND METHODS

2.1 The Initial Dataset

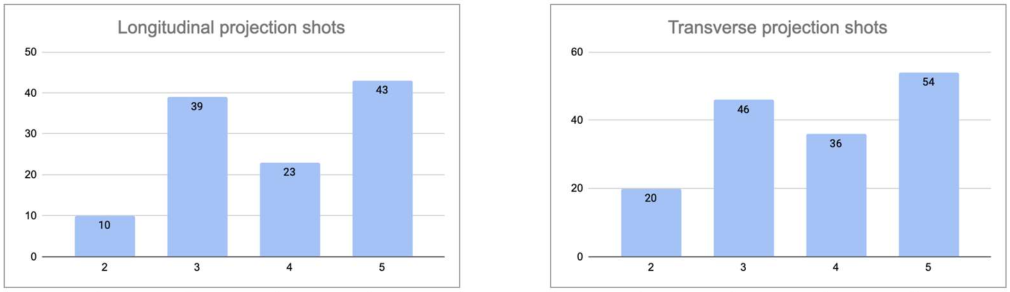
The initial dataset provided for training, validation, and testing of deep neural network models consisted of thyroid gland ultrasound examination tif cine loops of 80 patients in longitudinal (long) and transverse (cross) sections, marked masks, and class labels. The number of cine loops for these patients was more than one. The cine loops included from several dozen to several hundred frames. The selection of images of nodular formations was carried out at the Endocrinology Center in Moscow as part of project No. 22-15-00135 of the grant of the Russian Science Foundation.

The preprocessing of the initial dataset consisted of sequentially performed operations:

1) converting thyroid ultrasound files and masks from tif format to png images;
2) deleting text information;
3) removing black irrelevant areas;
4) bringing images to shades of gray;
5) resizing images and masks;
6) normalization of images.

When converting from tif to png, the tif components of the image were taken with a certain increment due to the high similarity of neighboring images.

The quantitative affiliation of nodal formations in the available dataset to the EU-TIRADS classes was analyzed (Figure 1).



EU-TIRADS	2	3	4	5	Total	EU-TIRADS	2	3	4	5	Total
Number	10 8,70%	39 33,91%	23 20,00%	43 37,39%	115 100,00%	Number	20 12,82%	46 29,49%	36 23,08%	54 34,62%	156 100,00%

Figure 1: Quantitative Affiliation of Nodal Formations in the Available Dataset to the EU-TIRADS Classes

2.2 Research of Approaches to Increasing the Amount of Data for Training Models

2.2.1 Experiment to test the effect of the number of used ultrasound image cine loops on the quality of trained models

The issue of the amount of data for training is especially relevant at the moment. Although the main task in designing new architectures is to optimize the structure of models in such a way that, with less volume and quality of data, more accurate predictions are obtained during validation and testing, many deep architectures still require a large amount of unique data for training. This experiment was based on the hypothesis that the volume of the training sample could not be increased by repeating similar images from the ultrasound of the cine loop of one patient and could be increased only by expanding the dataset with new unique copies of other patients and/or, in extreme cases, data from the augmentation process.

Let us consider the formulation of the detection problem. Let X be the set of feature descriptions of objects. The mapping $f: X \rightarrow D_f$is a feature f of object a. Here $D_f$is a set of acceptable values of the attribute. In the detection method under consideration, object a is an element of splitting the image into cells of a certain scale.

Each image of the object $a \in X$can be characterized by the values of features$f_i, i = 1, \dots, r$, the sets of which are the same for all objects. Therefore, the feature vector of the object $a \in X$can be determined by $x = \left(f_i(a), \dots, f_r(a)\right)$. Here the feature vector can be identified with the objects themselves.

To implement the detection task, it is necessary to construct the function $F: X\underset{\rightarrow}{}Y$that will map class $y_i \in Y$ to an arbitrary object from the X set with a certain probability from the predetermined probability distribution space. Here $Y = \{y_1, y_2, \dots, y_k,\}$is a finite set of classes, the partition into which exists on the entire set X [8].

Next, it is necessary to consider the formulation of the segmentation problem. The problem of semantic segmentation (pixel classification of images) can be considered as the problem of finding the evaluation function $h: X \rightarrow Y$for each pixel from the input image space X to the label space Y. The label space can include semantic maps or classification tags.

Considering the images $(x, y) \in X \times Y$ marked up into classes, it can be assumed that they belong to a fixed unknown probability distribution Ddefined on $X \times Y$.

Thus, the problem of finding h is reduced to finding the best indicator from a predefined functional space H(class of hypotheses) that restricts h and is selected based on knowledge about the segmentation problem being solved [9].

The present study was conducted on the example of two neural network architectures: YOLOv5 [10] to solve the detection problem and DeepLabV3 [11] to solve the segmentation problem.

15 datasets were used to train YOLOv5 architecture networks. The datasets were divided into three main categories: transverse images, longitudinal images, and all images. There were ашму sets of each category, where the following images were used: all images, every 3rd, every 5th, every 10th, and every 15th image. For validation, the sets were divided only into longitudinal, transverse, and all images. Figures 2 and 3 show the quantitative distribution of images for each of the sets.

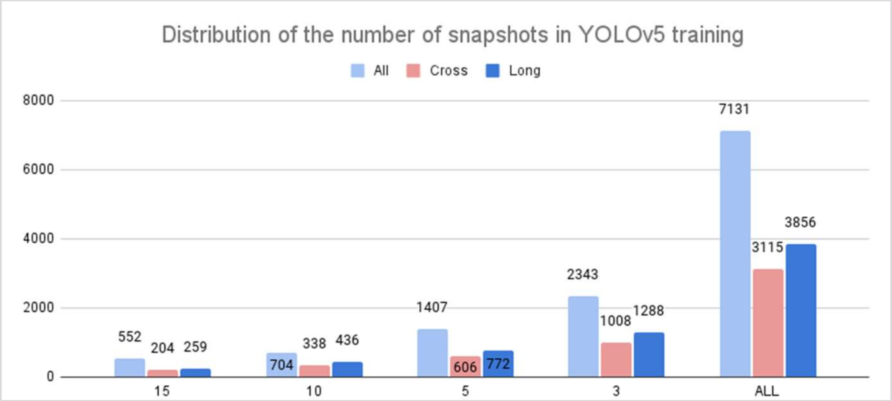


Figure 2: Distribution of the Number of Images during YOLOv5 Training

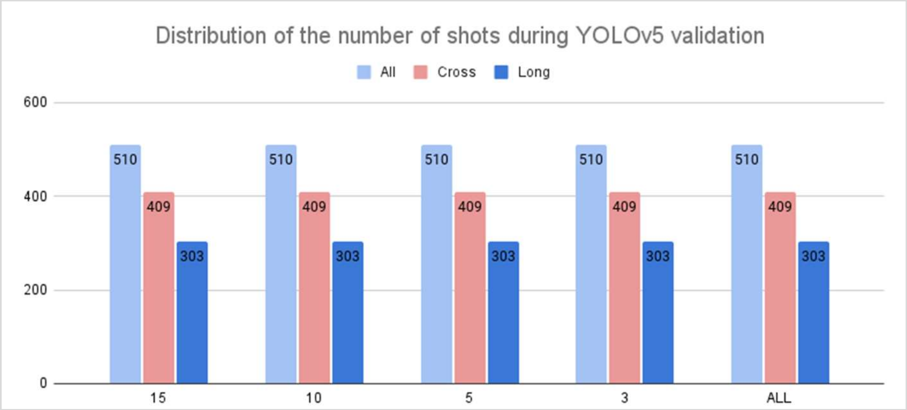


Figure 3: Distribution of the Number of Images for YOLOv5 Validation

To train DeepLabV3 networks, 6 sets of images were compiled. Each set included both transverse and longitudinal ultrasound projections. Each of the 6 sets included: all images, every 2nd, every 4th, every 8th, every 16th, and every 32nd image. Figure 4 shows the quantitative distribution of images for each of the sets. The colors of the chart columns are a symbol for interpreting learning outcomes. The models were trained on one number of epochs and validated on one set of images.

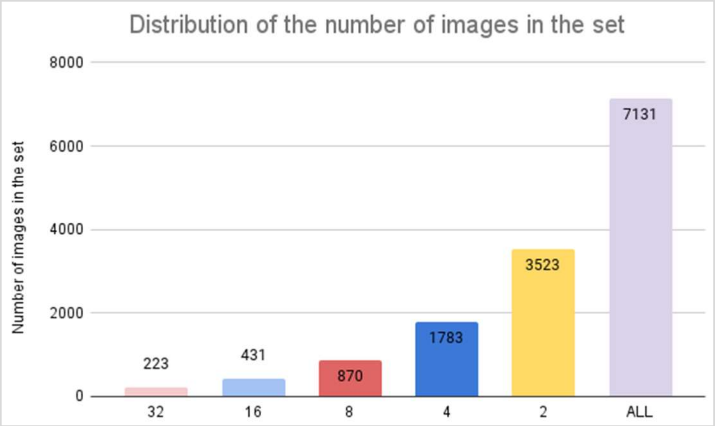


Figure 4: Distribution of the Number of Images during Training

According to the results of the training, the metrics of which are shown in Figures 5 and 6, the number of images does not affect the results of neural network learning. The differences in metrics are less than 10%. When training on each 32nd image, the Loss function begins to converge more slowly during testing. However, due to the small amount of data during training, the model is retrained faster. The networks showed the best result when training on transverse projections of thyroid ultrasound.

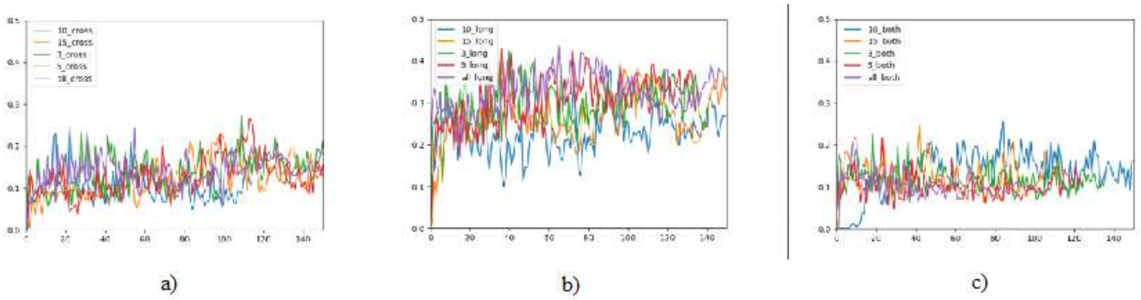


Figure 5: Indicators of the mAP-0.5 Metric in the Process of Training Models: a) Models with Training on Longitudinal Images; b) Models with Training on Transverse Images; c) Models with Training on All Images

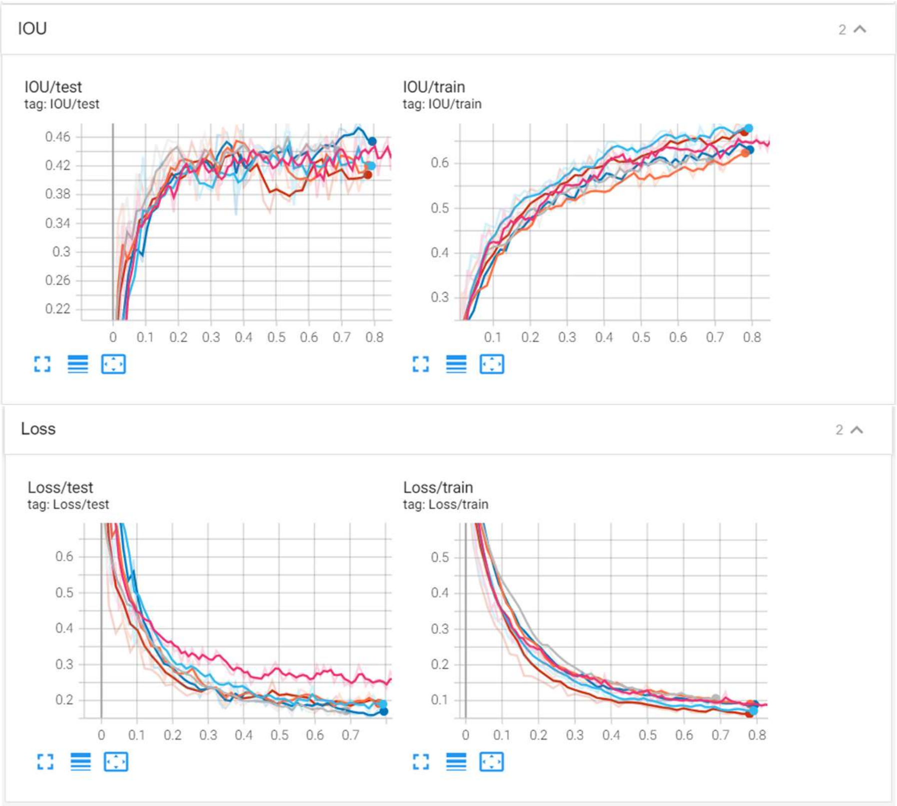


Figure 6: Indicators of Loss and Intersection over Union (IoU) Metrics during the Training and Testing of the Models: a) Loss and IoU during the Testing of the Models; b) Loss and IoU during the Training of the Models

The models obtained as a result of training were tested on the images of the test sample. The best test results were obtained using models trained on images of transverse ultrasound projections. The best indicators regarding the amount of data used in training were obtained following the results of validation, where the differences in indicator metrics varied within 10-15%.

Based on the indicators obtained during the validation and testing of models, as well as during the analysis of metrics during training, it was found that the quality of training was practically not affected by an overabundance of the same type of data. When training such deep architectures, it is necessary to have a large volume of unique images.

Summing up, we can conclude that similar images (images of the same projection of one patient) cannot be used as independent data. To improve the quality indicators of deep architectures, it is necessary to expand the training sample by increasing the number of images of different patients or using other methods to expand the dataset.

One of the ways to solve the described problem is to use the methods of augmentation of the training dataset with new synthesized images. The application of this approach is considered in the following experiment.

2.2.2 An experiment to test the effect of the image set augmentation on the quality of trained models

Formally, the task of augmentation of a set of images looks like this. There is an initial labeled set of N medical images X={x_1, x_2, …, x_N} with masks Y={y_1, y_2, …, y_N}. This set is divided into training and test samples, for example, in the ratio of 80% and 20%, respectively.

We consider the training of a neural network model M used to solve the problem of semantic image segmentation [12], with training parameters P for N_e epochs on a set of images X with masks Y. The segmentation quality of the trained model is evaluated by some Mt quality metric on a test sample of X_{test} images with Y_{test} masks. Let us denote the best quality of model segmentation in the test for the trained network as:

$$Q(M(P,N_e),X,Y) = min(Mt_i) \qquad (1)$$

where Mt_i is the value of the segmentation quality metric of the model on the test at the ith epoch (i is a natural number).

It is necessary to specify such a set of augmentation methods F that

$$Q(M(P,N_e),F[X],Y) > Q(M(P,N_e),X,Y) \qquad (2)$$

i. e., it is necessary to form such a set of augmentation methods for medical image sets that would improve the value of the chosen metric of the quality of solving the semantic segmentation problem by the model.

Based on the analysis of existing approaches to image augmentation [13-21], image augmentation methods are classified according to several criteria. According to the type of changes made to datasets, one can distinguish methods of geometric transformations (among which affine transformations are often mentioned), methods of transformations at the pixel level, and methods of creating artificial data using generative-adversarial neural networks.

The set of transformations applied to the training and test samples often differ. To generalize forecasts to test data, test-time augmentation (TTA) transformations are also used, the essence of which is to perform several different modifications for each image [14, 16, 17].

According to the frequency of use, transformations can be divided into constant ones, i. e. those that apply to all images of a given set, and non-constant ones which are applied with some probability or randomly from a given set of transformations [18].

Examples of the application of image geometric transformations are shown in Figure 7. Examples of using pixel-level transformation methods are shown in Figure 8.

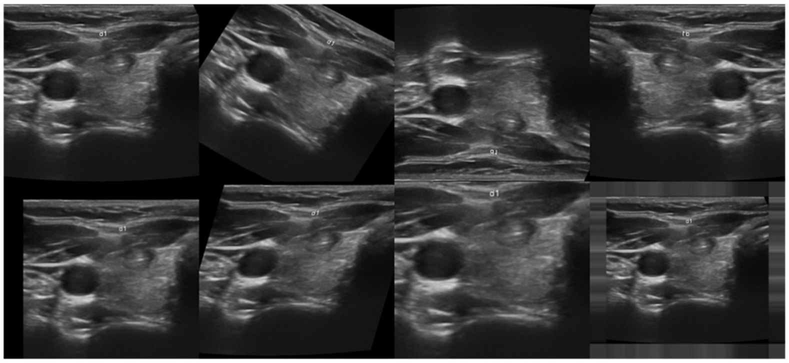


Figure 7: Geometric Methods of Image Augmentation
From left to right line by line: original image, rotation, mirror image along the vertical axis, mirror image along the horizontal axis, transfer, shift, zoom, crop

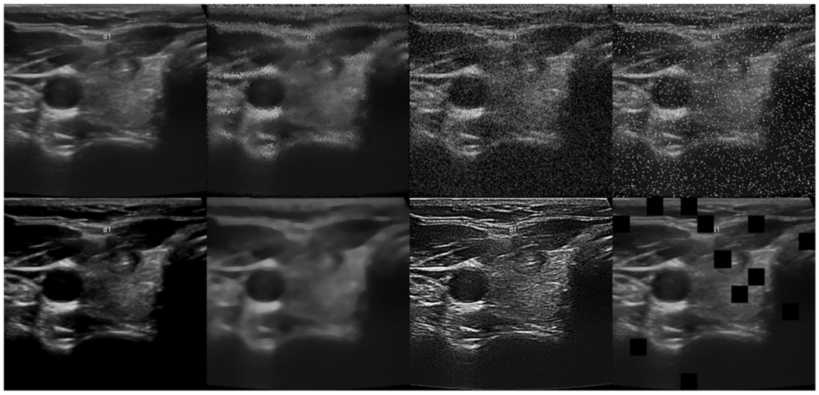


Figure 8: Image Augmentation Methods through Pixel-Level Transformations
From left to right line by line: original image, elastic transformation, Gaussian noise, salt-and-pepper noise, linear contrast, median filter, sharpen filter, dropout transformation

To solve the problem of semantic image segmentation, two networks with the same encoder-decoder structure were used, a detailed description of the sequential application approach of which is presented in [22].

To assess the quality of segmentation, IoU metrics and Dice coefficient (DC) were used. The average values of the corresponding metrics were calculated for several images.

Based on the analysis of approaches to expanding datasets, an experiment was set up to augment the existing set of thyroid ultrasound images, the results of which are presented in section 3.

Analysis of the results of the experiment allows us to conclude that the set of augmentation methods proposed for use has brought diversity to the initial data, improving the indicators of segmentation quality metrics and increasing the generalizing abilities of models that received ultrasound images in their entirety. However, the complication of the augmentation process of image sets containing enlarged areas with neoplasms harms the quality indicators of segmentation for models trained on such data.

3. RESULTS

3.1 Solving Classification and Segmentation Problems, Conducting Experiments to Assess the Quality of the Solutions Obtained

3.1.1 Solving the problem of semantic image segmentation

To find nodular neoplasms on thyroid ultrasound images with their boundaries highlighted, we solved the problem of semantic image segmentation, i. e. dividing images into segments (groups of pixels) and determining the type to which each segment belongs.

The study analyzed the existing approaches and architectures of deep neural networks to solve the problem of semantic segmentation. One of those is U-Net, an architecture created for the segmentation of biomedical images in the Computer Science department of the University of Freiburg. A newer model for semantic segmentation of 2D images, DeepLabV3+, was also studied and trained [23].

The experiment with the augmentation of datasets was carried out while training the DeepLabV3+ network with the EfficientNet-B6 encoder. To solve the problem of semantic segmentation, this architecture was used sequentially 2 times. The first segmentation network (hereinafter referred to as network 1) provided a rough localization of thyroid nodules in images with a size of 256x256 pixels. The second network (hereinafter referred to as network 2) had to segment the image more precisely, i. e. determine the position of the node on the region of interest (ROI) with a size of 512x512 pixels with a roughly localized node.

The available dataset was divided into training (80%) and test (20%) samples.

There were three experimental sets for augmentation methods: empty (without augmentation), simple, and complex ones. The simple set included the basic methods of geometric transformations, such as rotation, mirroring along the horizontal and/or vertical axis, transfer, shift, scaling, and cropping. The complex augmentation set provided for the transformation of not only the training but also the validation (test) sample. Augmentation of the training sample included geometric transformations of the entire image (rotation, mirroring along the horizontal and/or vertical axis, transfer, shift, scaling, and cropping), local affine transformation (scaling), and transformations at the pixel level (elastic transformations, Gaussian noise, linear contrast, sharpness change, Gaussian blur, average blur, median filter, and pixel zeroing). The described transformations were not applied all at the same time but were used with a given probability, in random order, to choose from one or more methods for the same type. The augmentation of the test set of images was a TTA including rotation and mirror reflections along the horizontal and vertical axes.

DeepLabV3+ networks were trained separately for networks 1 and 2, on longitudinal, transverse, and combined images simultaneously, without augmentation, with simple augmentation, and with complex augmentation. The total number of trained models was 18.

As a result of the experiment, the following results were obtained. The complication of augmentation of the training set of images with equal training parameters led to an increase in the values of the Loss function at the training stage, which indicates that the transformation methods bring variety to the original dataset. For network 1, the input of which received full images, the complexity of augmentation of the training and test sets improved the indicators of segmentation quality metrics, which indicates an increase in the generalizing abilities of the models. However, for network 2, the input of which received ROI, where the node occupied most of the image, the complexity of augmentation had a negative effect. These conclusions can be drawn from the analysis of the graphs of Figures 9 and 10.

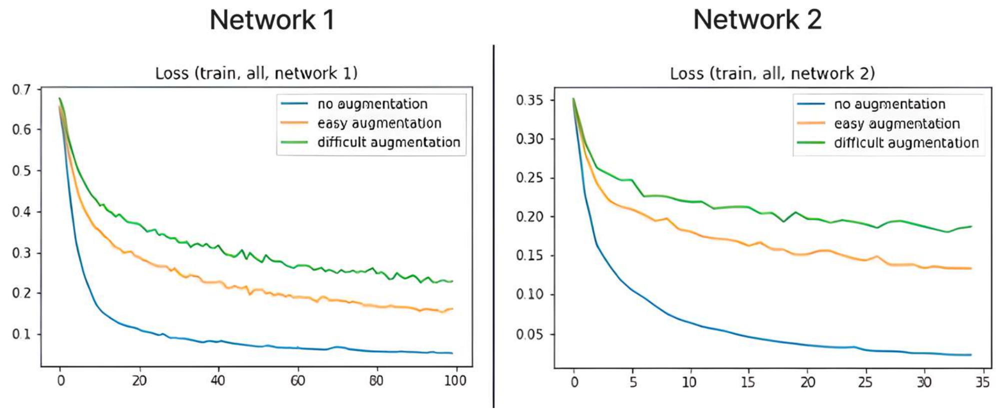


Figure 9: Graphs of the Values of the Loss Function during Training

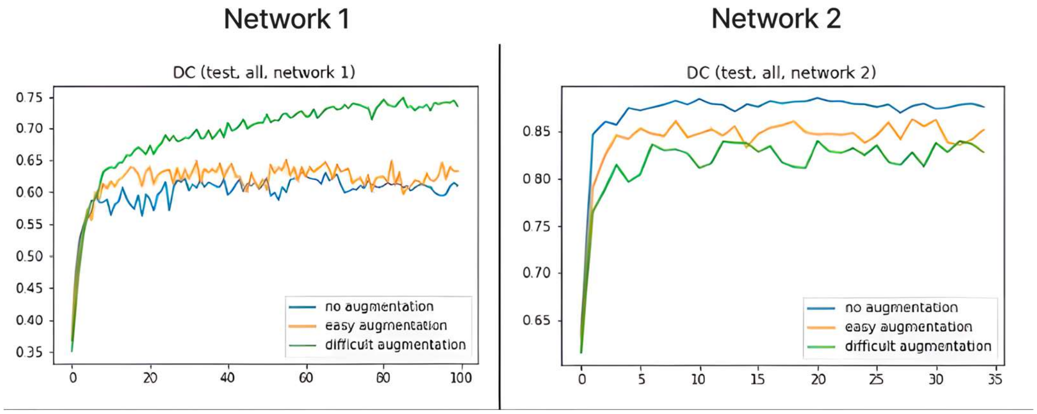
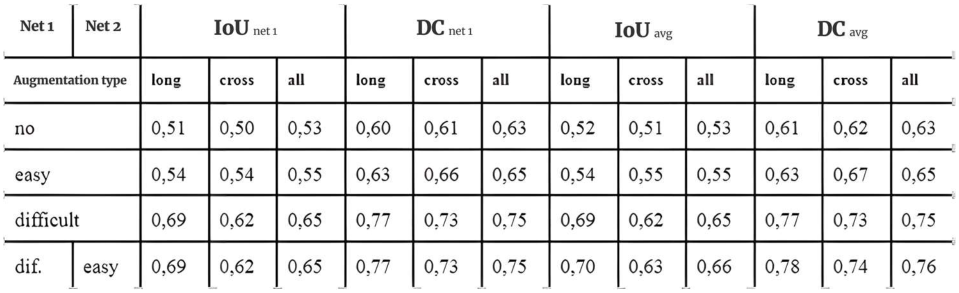


Figure 10: Graphs of DC Values on the Test

Another result of the experiment was that the consistent use of two networks had little effect on the quality of segmentation: either it did not change the indicators of quality metrics, or it improved them by only 1%. Figure 11 shows the values of segmentation quality metrics calculated based on the results of the sequential application of two networks (IoU_{av}, DC_{av}) in comparison with the values of segmentation quality metrics of only the first network ($IoU_{network\ 1}$, $DC_{network\ 1}$).



Net 1	Net 2	$IoU_{net 1}$			$DC_{net 1}$			IoU_{avg}			DC_{avg}		
Augmentation type		long	cross	all	long	cross	all	long	cross	all	long	cross	all
no		0,51	0,50	0,53	0,60	0,61	0,63	0,52	0,51	0,53	0,61	0,62	0,63
easy		0,54	0,54	0,55	0,63	0,66	0,65	0,54	0,55	0,55	0,63	0,67	0,65
difficult		0,69	0,62	0,65	0,77	0,73	0,75	0,69	0,62	0,65	0,77	0,73	0,75
dif.	easy	0,69	0,62	0,65	0,77	0,73	0,75	0,70	0,63	0,66	0,78	0,74	0,76

Figure 11: Values of Segmentation Quality Metrics of the First Network and Sequentially Two Networks

The best segmentation indicators on the test set of images were DC_{av}=83% and IoU_{av}=75%. Examples of good results and results worse than semantic ultrasound image segmentation are shown in Figure 12.

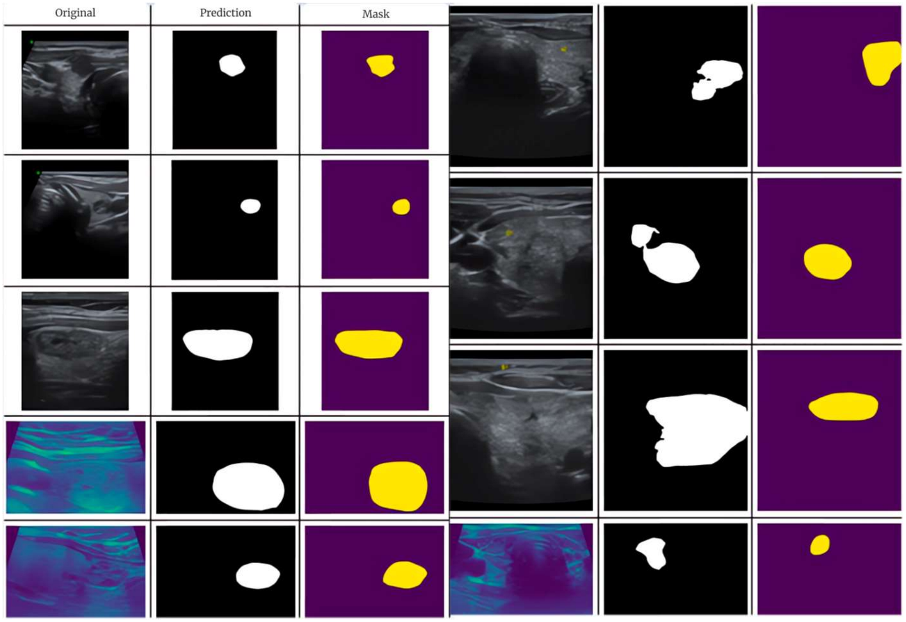


Figure 12: Examples of Semantic Segmentation Results by the DeepLabV3+ Network

Thus, as a result of working with semantic image segmentation networks, models solving this problem were trained, an experiment was conducted on augmentation of the original dataset and many image augmentation methods were formed that showed effectiveness and positively influenced the generalizing abilities of models when solving the problem of semantic image segmentation.

3.1.2 Solving the problem of image classification

To solve the problem of classifying images using the TIRADS classifier (6 categories), various deep learning structures with layer refinement were studied.

Due to the small number of ultrasound images with TIRADS-2 nodes in the original dataset, to reduce the imbalance between classes for training networks, images with TIRADS-2 and TIRADS-3 nodes were combined into one class.

Accuracy, precision, recall, and f1-score metrics were used to assess the classification quality of trained models:

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

$$precision = \frac{TP}{TP+FP} \tag{4}$$

$$recall = \frac{TP}{TP+FN} \tag{5}$$

$$f1score = \frac{2*recall*precision}{recall+precision} \tag{6}$$

Research and training were conducted for the ResNet architecture family (ResNet-18, ResNet-50, ResNet-101) [24, 25]. In the architectures of these networks, the first convolutional layer was changed (to supply 1-channel grayscale images, not 3-channel RGB images), as well as the last linear layer (according to the number of output classes equal to 3). The best indicators of classification metrics in the test sample were accuracy (60%), precision, recall, and f1-measure (from 58% to 60%) (achieved on ResNet-101). The low values of classification metrics are explained by the small amount of training data and the fact that although the models were pre-trained, this pre-training was not based on a set of medical images of a particular clinic.

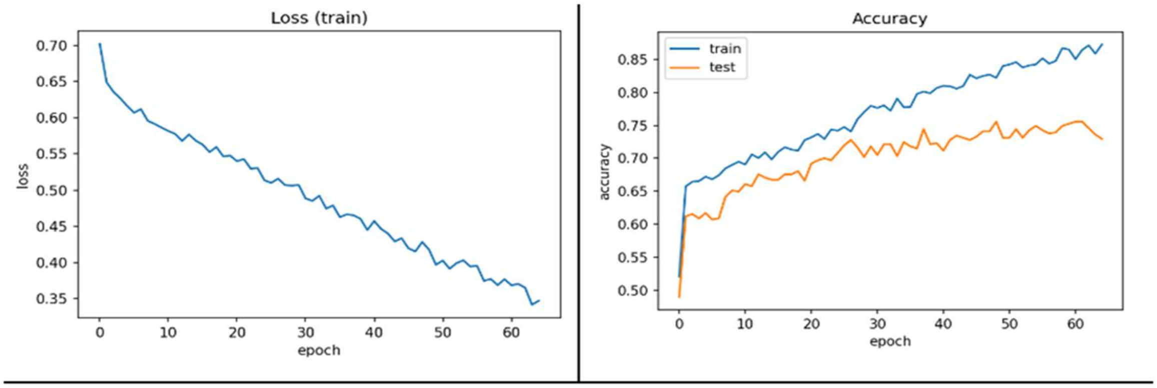
To improve the classification quality of a model from the EfficientNet architecture family trained on a set of small-volume images, it was decided to conduct training on a certain set of thyroid ultrasound images from open Internet sources, and then carry out additional training of some layers on the dataset available in the clinic.

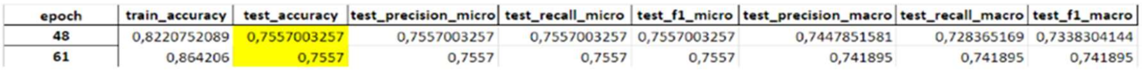
A dataset consisting of 3,493 thyroid ultrasound images was found in open sources to solve the problems of semantic image segmentation and classification, divided into two classes (benign and malignant formations). The dataset had been adjusted to the desired data markup format. The training sample included 2,879 images (82%), and the test sample had 614 images (18%). Quantitative belonging of images to classes by samples: in the training sample class 0 included 1,905 images (66%) and class 1 included 974 images (34%), while in the test sample class 0 included 378 images (62%) and class 1 included 236 images (38%).

The characteristics of the dataset (mean and standard deviation) were calculated.

To train image classification, we chose a model from a new family (EfficientNet) [26, 27], namely EfficientNet-B4, due to the relatively small number of parameters comparable to the number of parameters of the ResNet-50 network, but with a fairly high accuracy index on the ImageNet set.

In the architecture of the classification network, the first convolutional layer (for grayscale images) and the last linear layer (for the number of output classes, in this case, equal to two) were changed. The best accuracy score on the test sample (75.57%) was achieved by a model trained with the Adam optimizer using a scheduler to change the learning rate during training on the training sample, to which the complex augmentation described earlier was applied. The graph of the values of the Loss function in training, the graph of the accuracy metric values in training and test, and the best values of the classification metrics in the test are shown in Figure 13.





epoch	train_accuracy	test_accuracy	test_precision_micro	test_recall_micro	test_f1_micro	test_precision_macro	test_recall_macro	test_f1_macro
48	0,8220752089	0,7557003257	0,7557003257	0,7557003257	0,7557003257	0,7447851581	0,728365169	0,7338304144
61	0,864206	0,7557	0,7557	0,7557	0,7557	0,741895	0,741895	0,741895

Figure 13: A Graph of the Values of the Loss Function in Training, a Graph of the Accuracy Metric Values in Training and the Test, and the Best Values of Classification Metrics in the Test

However, retraining only the last layers on the existing dataset (with a large number of classes) did not lead to an increase in the values of classification quality metrics on the test above 60%, which can be explained again by a small number of training datasets and an increase in the number of classes of the pre-trained model. Then all layers of models were trained.

To identify the impact on the training of the nature of the dataset (full images or images from ROIs), models with the EfficientNet-B4 architecture were trained with the changes described earlier, on a dataset from the Internet from complete images without augmentation (M1) and with complex augmentation (M2), on a dataset from ROIs without augmentation (M3) and with complex augmentation (M4). The graph of accuracy metric values on the test of trained models is shown in Figure 14.

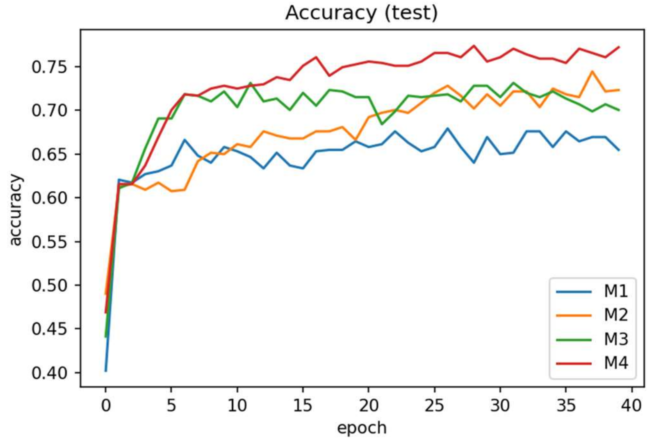
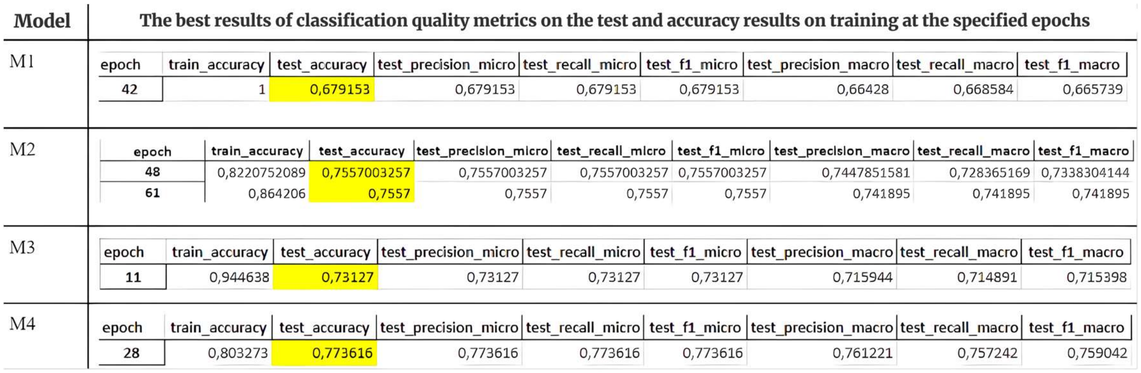


Figure 14: Graph of Accuracy Metric Values on a Test of Trained Models

The best results of classification quality metrics on the test for four trained models are shown in Figure 15.



Model	The best results of classification quality metrics on the test and accuracy results on training at the specified epochs								
	epoch	train_accuracy	test_accuracy	test_precision_micro	test_recall_micro	test_f1_micro	test_precision_macro	test_recall_macro	test_f1_macro
M1	42	1	0,679153	0,679153	0,679153	0,679153	0,66428	0,668584	0,665739
M2	48	0,8220752089	0,7557003257	0,7557003257	0,7557003257	0,7557003257	0,7447851581	0,728365169	0,7338304144
	61	0,864206	0,7557	0,7557	0,7557	0,7557	0,741895	0,741895	0,741895
M3	11	0,944638	0,73127	0,73127	0,73127	0,73127	0,715944	0,714891	0,715398
M4	28	0,803273	0,773616	0,773616	0,773616	0,773616	0,761221	0,757242	0,759042

Figure 15: The Best Indicators of Classification Quality Metrics on the Test and Accuracy Indicators on Training at the Specified Epochs

On ROIs and full images without augmentation, the models were quickly retrained: the best result on the test of the model trained on full images has an accuracy of 67.92%, and the model trained on ROIs has an accuracy of 73.13%. The best accuracy score on the test of a model trained on full images was 75.57%. The best accuracy score on the test of a model trained on ROIs with complex augmentation was 77.36%. Thus, the complex augmentation of a dataset consisting of ROIs provides higher model classification rates.

To determine a model capable of achieving higher classification indicators and trained on the ROIs of an existing small dataset, models with the architectures EfficientNet-B2 (M5), EfficientNet-B4 (M6), and EfficientNet-B6 (M7) with the changes described earlier were trained. The number of trained network parameters: 7.7 million, 17.6 million, and 40.7 million, respectively. The graph of accuracy metric values on the test of trained models is shown in Figure 16.

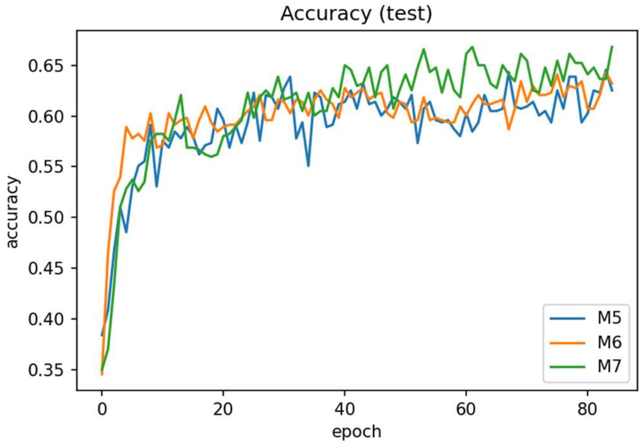
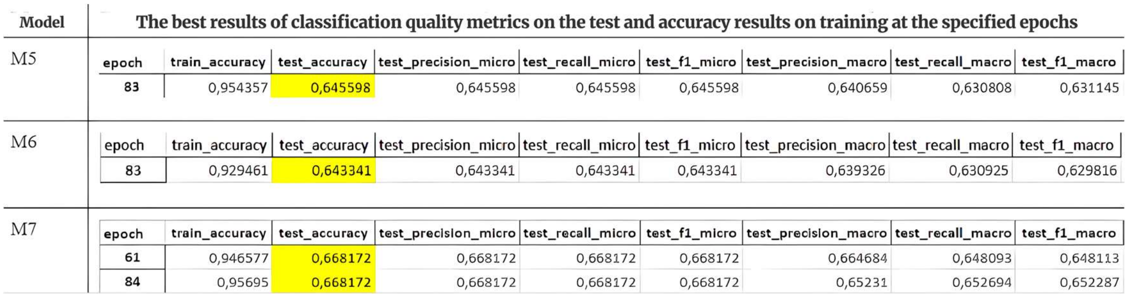


Figure 16: Graph of Accuracy Metric Values on a Test of Trained Models

The best results of classification quality metrics on the test for three trained models are shown in Figure 17.



Model	epoch	train_accuracy	test_accuracy	test_precision_micro	test_recall_micro	test_f1_micro	test_precision_macro	test_recall_macro	test_f1_macro
M5	83	0,954357	0,645598	0,645598	0,645598	0,645598	0,640659	0,630808	0,631145
M6	83	0,929461	0,643341	0,643341	0,643341	0,643341	0,639326	0,630925	0,629816
M7	61	0,946577	0,668172	0,668172	0,668172	0,668172	0,664684	0,648093	0,648113
	84	0,95695	0,668172	0,668172	0,668172	0,668172	0,65231	0,652694	0,652287

The best results of classification quality metrics on the test and accuracy results on training at the specified epochs

Figure 17: The Best Indicators of Classification Quality Metrics on the Test and Accuracy Indicators on Training at the Specified Epochs

It can be seen that the models on the test had fairly similar classification results, but the EfficientNet-B6 model with the changes made to the architecture trained on a set of small-volume ROIs images achieved higher accuracy = 66.82%.

The effect of an increase in the volume of the dataset (+25%) on the indicators of the trained network can be traced in Figures 18 and 19 (the EfficientNet-B6 network, as amended). Before the dataset was increased, the best accuracy score on the test was 66.82%, after the increase it was 70.75%.

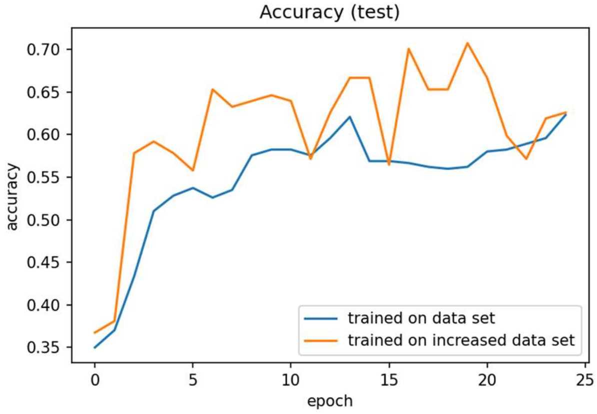
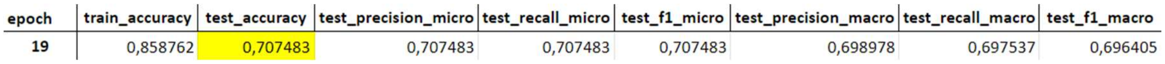


Figure 18: Graph of Accuracy Metric Values on the Test of Trained Models before and after Increasing the Dataset



epoch	train_accuracy	test_accuracy	test_precision_micro	test_recall_micro	test_f1_micro	test_precision_macro	test_recall_macro	test_f1_macro
19	0,858762	0,707483	0,707483	0,707483	0,707483	0,698978	0,697537	0,696405

Figure 19: The Best Indicators of Classification Quality Metrics on the Test and Accuracy Indicators on Training at the Specified Epoch after Increasing the Dataset

Thus, as a result of working with image classification networks, various architectures were investigated, some ways of solving the problem of a small amount of data for training were tested, the effects of image augmentation in solving the classification problem, the nature of the original dataset (full images or images from ROIs) and the volume of the training sample on the abilities of the model were evaluated.

4. DISCUSSION

The issue of the amount of data for training is most relevant at the moment. Although the main task in designing new architectures is to optimize the structure of models in such a way that with the least possible amount and quality of data, the most accurate predictions are obtained during validation and testing, many deep architectures require a large amount of unique data. From the results obtained in this study, it follows that the volume of the training sample should be increased not by repeating similar images of one patient, but by expanding the dataset with new unique instances of other patients and/or data from the augmentation process. The data obtained have been confirmed or refuted in other works on other datasets.

For example, the results of the study [9] confirm our conclusions made during this study. This paper examines the impact of data volume on the training of a deep electrocardiogram (ECG) classification architecture. The paper analyzes the use of different amounts of non-augmented data during network training. It should be noted that the training sample consisted only of unique ECG images. As a result of the described experiment, it is concluded that the metrics of network learning evaluation reach a plateau when using 45% of the training sample. In addition, an experiment to analyze the effect of the volume of one instance of a training sample demonstrates that the size of a unique training instance does not affect the learning outcomes. After reaching the optimal volume for training, the metrics reach a plateau.

The paper [28] discusses experiments to assess the impact of the quality and volume of data for training models. Deep neural architectures and various medical data samples are considered here. The problems found as a result of the study confirm our conclusions about the need for high-quality images and a large amount of unique data.

In [28], the dependence of the prediction accuracy of neural network architectures on the volume and quality of the training sample is investigated. This study rather proves that a large amount of data is the most important aspect in the formation of hyperparameters of the model. This conclusion also coincides with the results obtained in this study. In the described study, work is carried out with both unique and augmented data instances (LVIS-OneShot, PASCAL-5, FSS-1000), as in the experiments of this work.

The paper [29] describes the methods of medical image augmentation for the images which are used for the diagnosis, treatment, and monitoring of various diseases. The main focus is on the latest achievements in the field of medical tomography, which allows for obtaining images of internal organs and tissues in high resolution. The paper presents examples of the use of medical images in the diagnosis of oncological diseases, neurological disorders, cardiovascular diseases, and other pathologies. The results of the experiment showed that the use of data augmentation could significantly improve the accuracy of object recognition in images. In addition, the increase in the training sample also achieved an improvement in the accuracy of object recognition, which confirms the conclusions obtained in this study.

In contrast to our study of nodular neoplasms of the thyroid gland, the paper [30] focused on the brain diseases. The authors investigated the use of deep learning for the automatic classification of brain diseases based on magnetic resonance imaging (MRI) images. They used the DeepLabV3+ neural network to detect and classify diseases and compared its work with other classification methods. The results showed that the use of the DeepLabV3+ architecture for the classification of brain diseases was an effective method and a promising area for further research in the field of medical image analysis.

5. CONCLUSION

In this paper, the use of neural network architectures for solving segmentation, classification, and detection problems was investigated. To do this, several experiments were conducted to create and train deep architectures for various types of analysis.

In the course of the study, the influence of the volume and quality of data for training deep architectures was considered, and we analyzed the effectiveness of using various methods of the medical image set augmentation in a training sample of neural networks in solving semantic segmentation, classification, and detection problems. To carry out the analysis, the thyroid ultrasound scans of 166 patients with nodular formations were used. The selection of patients and the recording of a cine loop of nodular formations was carried out within the framework of project No. 22-15-00135 funded by a grant from the Russian Science Foundation.

In the course of the work, we performed a comparative analysis of the results of training the model with different types and images and different amounts of data. When training and testing models on a large number of the same type of data, metric indicators showed a spread of only 10%, which is not significant for the architectures under study, but it was experimentally proved that a set of augmentation methods introduced diversity into the source data, improving the metrics of segmentation quality and increasing the generalizing abilities of models that received full ultrasound images.

In the course of all the experiments conducted, models were obtained and approaches were identified for working with medical images of ultrasound examinations of the thyroid gland. It was shown that to improve the quality of deep architectures, it was necessary to expand the training sample by increasing the number of images of different patients and using augmentation methods. In the course of the study, neural network architectures were identified that were most suitable for working with medical images, and models with high accuracy were obtained, both at the training stage and the testing stage. In further research, it is necessary to continue studying the use of various artificial intelligence methods for personalized medicine.

REFERENCES:

[1] P.J. Brockwell, and R.A. Davis, *Introduction to time series and forecasting*. Springer, Cham, 2016.

[2] A. Aldweesh, A. Derhab, and A.Z. Emam, “Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues,” *Knowledge-Based Systems*, Vol. 189, 2020, Art. No. 105124. http://dx.doi.org/10.1016/j.knosys.2019.105124

[3] J. Chen, H. You, and K. Li, “A review of thyroid gland segmentation and thyroid nodule segmentation methods for medical ultrasound images”, *Computer Methods and Programs in Biomedicine*, Vol. 185, 2020, Art. No. 105329. https://doi.org/10.1016/j.cmpb.2020.105329

[4] P. Deng, X. Han, X. Wei, and L. Chang, “Automatic classification of thyroid nodules in ultrasound images using a multi-task attention network guided by clinical knowledge”, *Computer Methods and Programs in Biomedicine*, Vol. 150, 2022, Art. No. 105329. https://doi.org/10.1016/j.compbiomed.2022.106172

[5] X. Zhang, V.C. Lee, J. Rong, J.C. Lee, and F. Liu, “Deep convolutional neural networks in thyroid disease detection: A multi-classification comparison by ultrasonography and computed tomography”, *Computer Methods and Programs in Biomedicine*, Vol. 220, 2022, Art. No. 106823. https://doi.org/10.1016/j.cmpb.2022.106823

[6] Y. Sharifi, M.A. Bakhshali, T. Dehghani, M. Danaiashgzari, M. Sargolzaei, and S. Eslami, “Deep learning on ultrasound images of thyroid nodules”, *Biocybernetics and Biomedical Engineering*, Vol. 41, No. 2, 2022, pp. 636-655. https://doi.org/10.1016/j.bbe.2021.02.008

[7] F.N. Tessler, W.D. Middleton, and E.G. Grant, “Thyroid imaging reporting and data system (TI-RADS): A user’s guide”, *Radiology*, Vol. 287, No. 1, 2018, pp. 29-36. https://doi.org/10.1148/radiol.2017171240

[8] T. Luddecke, and A. Ecker, “The role of data for one-shot semantic segmentation”, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021, pp. 2653-2658.

[9] M. Toldo, A. Maracani, U. Michieli, and P. Zanuttigh, “Unsupervised domain adaptation in semantic segmentation: A review”, *Technologies*, Vol. 8, No. 2, 2020, Art. No. 35. https://doi.org/10.3390/technologies8020035

[10] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You only look once: Unified, real-time object detection”, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 779-788. https://doi.org/10.1109/CVPR.2016.91

[11] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation”, 2017. https://doi.org/10.48550/arXiv.1706.05587

[12] S. Zhou, D. Nie, E. Adeli, Q. Wei, X. Ren, X. Liu, E. Zhu, J. Yin, Q. Wang, and D. Shen, “Semantic instance segmentation with discriminative deep supervision for medical images”, *Medical Image Analysis*, Vol. 82, 2022, Art. No. 102626.

[13] K. Maharana, S. Mondal, and B. Nemade, “A review: Data pre-processing and data augmentation techniques”, *Global Transitions Proceedings*, Vol. 3, No. 1, 2022, pp. 91-99. https://doi.org/10.1016/j.gltp.2022.04.020

[14] J. Nalepa, M. Marcinkiewicz, and M. Kawulok, “Data augmentation for brain-tumor segmentation: A review”, *Frontiers in Computational Neuroscience*, Vol. 13, 2019, Art. No. 83. https://doi.org/10.3389/fncom.2019.00083

[15] P. Chlap, H. Min, N. Vandenberg, J. Dowling, L. Holloway, and A. Haworth, “A review of medical image data augmentation techniques for deep learning applications”, *Journal of Medical Imaging and Radiation Oncology*, Vol. 65, No. 5, 2021, pp. 545-563. https://doi.org/10.1111/1754-9485.13261

[16] D. Hoar, P.Q. Lee, A. Guida, S. Patterson, C.V. Bowen, J. Merrimen, C. Wang, R. Rendon, S.D. Beyea, and S.E. Clarke, “Combined transfer learning and test-time augmentation improves convolutional neural network-based semantic segmentation of prostate cancer from multi-parametric MR images”, *Computer Methods and Programs in Biomedicine*, Vol. 210, 2021, Art. No. 106375. https://doi.org/10.1016/j.cmpb.2021.106375

[17] *Image Test Time Augmentation with PyTorch*. TTAch. [Online]. Available: https://github.com/qubvel/ttach (access date: March 1, 2023).

[18] *Documentation of the imgaug library for image augmentation*. [Online]. Available: https://imgaug.readthedocs.io/en/latest/ (access date: March 1, 2023).

[19] Z. Hussain, F. Gimenez, D. Yi, and D. Rubin, “Differential data augmentation techniques for medical imaging classification tasks”, in *AMIA Symposium*, Washington, DC, USA, 2017, pp. 979-984.

[20] Y. Chen, X.H. Yang, Z. Wei, A.A. Heidari, N. Zheng, Z. Li, H. Chen, H. Hu, Q. Zhou, and Q. Guan, “Generative adversarial networks in medical image augmentation: A review”, *Computers in Biology and Medicine*, Vol. 144, 2022, Art. No. 105382. https://doi.org/10.1016/j.compbiomed.2022.105382

[21] G. Shi, J. Wang, Y. Qiang, X. Yang, J. Zhao, R. Hao, W. Yang, Q. Du, and N.G.-F. Kazihise, “Knowledge-guided synthetic medical image adversarial augmentation for ultrasonography thyroid nodule classification”, *Computer Methods and Programs in Biomedicine*, Vol. 196, 2020, Art. No. 105611. https://doi.org/10.1016/j.cmpb.2020.105611

[22] M. Wang, C. Yuan, D. Wu, Y. Zeng, S. Zhong, and W. Qiu, “Automatic segmentation and classification of thyroid nodules in ultrasound images with convolutional neural networks”, in N. Shusharina, M.P. Heinrich, and R. Huang (Eds.), *Segmentation, classification, and registration of multi-modality medical imaging data. MICCAI 2020. Lecture notes in computer science*, Vol. 12587, pp. 109-115. Springer, Cham, 2021. http://dx.doi.org/10.1007/978-3-030-71827-5_14

[23] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation”, in V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss (Eds.), *Computer vision – ECCV 2018. ECCV 2018. Lecture notes in computer science*, Vol. 11211, pp. 833-851. Springer, Cham, 2018. https://doi.org/10.1007/978-3-030-01234-2_49

[24] *ResNet (34, 50, 101): "ostatochnye" CNN dlya klassifikatsii izobrazhenii* [the "residual" CNN for image classification], January 29, 2019. [Online]. Available: https://neurohive.io/ru/vidy-nejrosetej/resnet-34-50-101/ (access date: March 1, 2023).

[25] *ResNet.* PyTorch Documentation. [Online]. Available: https://pytorch.org/vision/stable/models/resnet.html (access date: March 1, 2023).

[26] M. Tan, and Q.V. Le, “EfficientNet: Rethinking model scaling for convolutional neural networks”, in *ICML 2019. Machine learning, computer vision, and pattern recognition*, Long Beach, California, USA, 2019. https://doi.org/10.48550/arXiv.1905.11946

[27] *EfficientNet*. PyTorch Documentation. [Online]. Available: https://pytorch.org/vision/stable/models/efficientnet.html (access date: March 1, 2023).

[28] A.R. Luca, T.F. Ursuleanu, L. Gheorghe, R. Grigorovici, S. Iancu, M. Hlusneac, and A. Grigorovici, "Impact of quality, type, and volume of data used by deep learning models in the analysis of medical images", *Informatics in Medicine Unlocked*, Vol. 29, 2022, Art. No. 100911. https://doi.org/10.1016/j.imu.2022.100911

[29] A. Kebaili, J. Lapuyade-Lahorgue, and S. Ruan, "Deep learning approaches for data augmentation in medical imaging: A review", *Journal of Imaging*, Vol. 9, No. 4, 2023, Art. No. 81. https://doi.org/10.3390/jimaging9040081

[30] H. Polat, "Multi-task semantic segmentation of CT images for COVID-19 infections using DeepLabV3+ based on dilated residual network", *Physical and Engineering Sciences in Medicine*, Vol. 45, 2022, pp. 443-455. https://doi.org/10.1007/s13246-022-01110-w