








Using Deep Learning to Generate and Classify Thyroid Cytopathology Reports According to The Bethesda System



Eugene Diuldin¹, Artem Makanov¹, Boris Shifman², Elizabeth Bobrova¹, Stanislav Osnovin¹, Konstantin Zaytsev^{1*}, Aleksander Garmash¹, Fatima Abdulkhabirova²

¹ Institute of Intelligent Cybernetic Systems, National Research Nuclear University MEPHI (Moscow Engineering Physics Institute), Moscow 115409, Russia

² Laboratory of Cytology and Cytogenetics of the Department of Fundamental Pathomorphology, National Medical Research Centre for Endocrinology, Moscow 117036, Russia

Corresponding Author Email: KSZajtsev@mephi.ru

Copyright: ©2024 The authors. This article is published by IETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ria.380237>

ABSTRACT

Received: 4 October 2023
Revised: 12 February 2024
Accepted: 12 March 2024
Available online: 24 April 2024

Keywords:

Bethesda, deep learning, embedding, fine-needle aspiration, nodular goiter, thyroid, thyroid nodules, transformer

The purpose of this paper is to study approaches to the intellectual processing of Russian-language textual medical information concerning the thyroid cytopathology description to solve the issues of their classification and generation of the text of the medical report, as well as augmentation of descriptions in their acute shortage. Over the past decade, the field of biomedicine has not undergone significant changes. Approaches to analyzing patients' problems are mostly based on manual processing and expert knowledge of doctors. The paper considers the creation of a machine-learning pipeline containing a full cycle of data preprocessing and model training in the field of thyroid nodules fine-needle aspiration classification according to the Bethesda thyroid cytopathology reporting system. Sequential and transformer neural networks were used to design the architecture of deep learning models. The paper proposes approaches for cleaning and preprocessing raw medical descriptions to the required type. The obtained results show that sequential neural networks have greater accuracy on small data sets, and transformation architectures are superior to others when generating cytopathological reports on large amounts of data. The solution obtained in the study can be used as an additional reference tool for thyroid cytologists.

1. INTRODUCTION

Attempts to integrate natural language processing (NLP) methods in areas where only classical approaches have existed for a long time become increasingly evident. The use of new approaches is justified by the fact that classical methods often do not give the desired results and require constant improvements. Therefore, modern natural language text processing technologies are being introduced increasingly often into areas not previously covered by them.

One of these areas is medicine, where natural language is used to record individual results. Having the fixed signs of the disease described by the doctor in natural language, one needs to solve the issues of classifying the disease according to the accepted scale and automatically generating the text of the medical report indicating all parameters used for subsequent treatment.

Let us consider the formulation and solution of the tasks of working with medical texts written in natural language in the case of classification and generation of a medical opinion on the description in the field of cytological studies of the thyroid gland.

The initial data are:

- The Bethesda System for Reporting Thyroid Cytopathology (TBSRTC) [1, 2] according to which a thyroid

nodule fine needle aspiration (FNA) material is classified into one of six diagnostic categories with different corresponding malignancy risk (ranging from 4 to 97%), that helps to determine patient further management tactics.

- A corpus of pairs of real texts (1. description and 2. Conclusion-a TBSRTC category) made by cytopathologists of the National Medical Research Center for Endocrinology. However, in this case, in many pairs, the first element (a description) is missing due to its placement in another field along with the diagnostic category.

It is necessary to build a sequence of mathematical models (pipeline) for the formation of a medical opinion according to the description indicating the Bethesda category. The results of the work may be of interest as a part of a thyroid cytopathology automated analysis system and can be used to generate a description and the corresponding category for a FNA sample.

The task in its general formulation in different fields of activity was solved by several research groups. Therefore, it is reasonable to consider the closest approaches to the classification and generation of text tokens and problems of time complexity, spatial storage, and processing of raw text arrays.

The related works are discussed further in section 2. Section 3 outlines the proposed approach to the automatic generation

of Bethesda categories. In section 4, the results of the conducted quality studies of the proposed approach are discussed, and in section 5, similar works are described. Finally, section 6 summarizes the results of the study with conclusions and possibilities for its development.

This study primarily aims to innovate in the field of medical text processing by developing deep learning models tailored for Russian-language thyroid cytopathology reports. We specifically focus on utilizing Sequential and Transformer neural networks to classify diseases and generate comprehensive medical reports. The scope of this research is carefully defined to include an in-depth analysis of these neural network models, particularly assessing their performance in handling datasets of varying sizes.

2. LITERATURE REVIEW

Currently, several review papers offer text classification algorithms or put forward interesting ideas for evaluating medical text corpora. Thus, in the research [3], the use of complex approaches based on autoencoders and the use of convolutional networks for the classification of medical texts is discussed. The idea of mixing the received and generated data during learning as a method of improving the quality of the final model is interesting. This is important for us since to increase the body of raw data of missing medical descriptions, we have to solve the task of augmenting them based on available medical reports.

In the research [4], which focuses on learning text classification algorithms from scratch and obtaining the best representation only on similar data, approaches to obtaining training results on raw data are described, as opposed to widespread approaches to retraining architectures like BERT (Bidirectional Encoder Representations from Transformers).

The authors of the papers [5-7] involve architectures based on sequential and transformer networks to solve analytical medical problems, which can be applied to our task for the transfer of knowledge in the cross-domain field.

One of the tasks of this work is the classification according to a cytopathology description text, i.e., the formation of the Bethesda category [8] using deep learning. Using new approaches for text preprocessing and model training, both canonical transformers and their modifications, we will try to increase the accuracy of predictions relative to a given target threshold.

The solution to this problem lies in the way of creating a pipeline for preprocessing data and extracting key information from raw medical text corpora. The pipeline model is both a classifier of the Bethesda target label and a generator of a medical opinion on the medical description of the patient's problems.

The structure of this study is presented in the form of a sequence of sections. The first section describes the approaches in the preprocessing of the source data. In the second, a basic statistical model is created, according to which more advanced architectures will be evaluated. The following sections describe sequential approaches aimed at producing a short conclusions (diagnostic categories) based on a cytopathological description. Next, the transformer technologies and their training options are discussed. In conclusion, the results are summed up, and the validation of the resulting models is performed.

In the realm of medical text processing, various techniques

have been explored in the literature. Sequential neural networks, known for their simplicity and effectiveness in small data scenarios, have been widely used for structured data analysis. They excel in tasks where data points are interdependent, such as time-series analysis. On the other hand, Transformer neural networks, a more recent development, are noted for their ability to handle larger datasets efficiently.

Their architecture, based on self-attention mechanisms, allows them to process data points in parallel, making them particularly suitable for complex tasks like language translation and large-scale text generation. The literature also highlights hybrid approaches, combining the strengths of different models to optimize performance. These comparisons underscore the importance of choosing the right technique based on the specific requirements of the dataset and the task at hand.

3. MATERIALS AND METHODS

3.1 Data preparation

Working with medical reports on the classification of thyroid nodules raises the task of marking and clearing data. The corpus of real data used in the study contains more than 27 thousand thyroid cytopathology reports gathered in the Laboratory of Cytology and Cytogenetics of the Department of Pathomorphology of the Center for Endocrinology over ten years of work (2013-2023). The data was received in a blinded form (without personal information).

Medical texts have a weak internal structure, which leads to problems of clearing data from unnecessary tokens and highlighting the basic information necessary for further generation of signs.

There may be several Bethesda labels in one medical description. For example, "p11-1: uninformative material: peripheral blood cells, thyrocytes were not detected in the smear against the background of a thick colloid (according to Bethesda Thyroid Classification, category I)"; further in the same description: "l11-2: a node of colloidal to varying degrees proliferating goiter with cystic changes has been punctured (according to Bethesda Thyroid Classification, category II)".

Solving problems in the order of allocation of the target Bethesda label and key description tokens on an untagged set of data, we place the main emphasis on the primary use of regular expressions and probabilistic search for the most frequently encountered sentence tokens. To search for this pattern, a regular expression of the type '[IV]+' is used, which helps to find all possible combinations of the Bethesda label in the considered text. Sentences from the original data corpus, after clearing the data with the generated labels, can be grouped by class, getting the percentage of occurrence (Table 1).

Table 1. Percentage occurrence of labels in the case

Class (Index)	Number of Class Entries	Bethesda Entry Percentage
1	3,409	12.43
2	17,799	69.42
3	1,050	3.83
4	2,160	7.88
5	1,076	3.92
6	952	3.47

In this table, the general target field is presented after splitting groups of complex text queries into basic ones (with one class label), which expands the original groups with a class zero, without affecting the distribution of Bethesda labels previously affixed by doctors.

Part of the process of preprocessing raw data to obtain a more voluminous object space is the search for the number of sentences in the text and the probability of words appearing in each sentence, which is a hyperparameter for various types of data augmentation and validation of the obtained medical assessment reports. At the preprocessing stage, long text conclusions of patients, who had gone through multiple FNA due to having a multinodular goitre and therefore had several conclusions (Bethesda diagnostic categories), were divided. When dividing the data into logical elements, where each element was a sentence with a Bethesda label, the feature field expanded by 7.5%.

Further, after separating the classifier labels by cytopathological descriptions and diagnostic categories, the necessary tokens were allocated.

After cleaning and creating the resulting dataset with separated Bethesda labels in the range from 1 to 6, we obtained completely separated data to solve the problems of classifying the Bethesda label and generating a diagnostic category based on the medical description of the thyroid problem. Examples of cytopathological descriptions are presented in Table 2.

Table 2. Examples of cytopathological descriptions and corresponding labels (Bethesda diagnostic categories)

Cytopathological Descriptions	Label
<i>The material is not informative enough: Single dystrophic thyrocytes were found in thick smears with an abundant admixture of erythrocytes</i>	1
<i>In the smears, the cytogram is characteristic of a colloidal, to varying degrees proliferating B-cell goiter with areas of adenomatosis and cystic degeneration</i>	2
<i>In the smear with a heterogeneous cellular composition, among the abundance of erythrocytes and colloid, there are both groups of polymorphic thyrocytes with signs of goiter transformation and clusters of enlarged epithelial cells of irregular shape, with, sparse chromatin, densely located in shapeless and papillary-like structures</i>	3
<i>In the smear of high cellularity with a large admixture of blood against the background of the contents of the cystic hemorrhagic cavity, there are clusters of large epithelial cells with a wide cytoplasm, eccentric nuclei, pronounced degenerative changes, forming predominantly mixed structures and located separately, more suspicious in terms of the follicular formation of the thyroid gland from B cells</i>	4
<i>The sample consists of complexes of atypical epithelial cells with single intracellular pseudoinclusions against the background of cystic hemorrhagic changes and lymphocytic inflammation</i>	5
<i>Numerous isolated groups of thyroid papillary cancer cells have been found in the smear, forming trabecular and papillary structures with single intra-row pseudoinclusions</i>	6

The preprocessed source data (the main keyword tokens and the cleaned raw data) were saved in .csv format.

Ethical expertise. The protocol of the study was reviewed and approved by the local Ethical Committee of the National Medical Research Center for Endocrinology (Protocol No. 14

dated 25.07.2023).

3.2 Creation of a statistical model

When building a complex system for classification according to Bethesda and generating a diagnostic category based on a cytopathological description, it is necessary to be sure that the model can be effective in terms of temporal and spatial complexity.

In the previously reviewed papers, CNN (convolutional neural networks) and LSTM (long short-term memory) networks with modifications of the original algorithm were used to solve these problems. Such approaches work well for the task but are difficult to interpret and also slow, which leads to time costs.

Basic models are suitable for estimating temporal and spatial capacities, which will be emphasized in the validation of the final architecture in the future.

To solve problems step by step, one wants to combine the embedding approach (a numerical vector derived from words) and clustering methods to determine the class label. This choice helps to optimize the query execution time [9]. In this study, we use FastText-based approaches [5] and methods for normalizing the final vector representations.

The first method we will consider uses document token norms and clustering using the K-means algorithm. This method allows us to quickly determine the Bethesda label. After normalization and training of the algorithm, we make predictions and compare them with the target labels of the test dataset.

The algorithm divides the set of vector space elements into a pre-known number of clusters k. The main idea is that at each iteration the center of mass is recalculated for each cluster obtained at the previous step, then the vectors are divided into clusters again in accordance with which of the new centers turned out to be closer according to the selected metric. The algorithm ends when at some iteration there is no change in the intra-cluster distance. This happens in a finite number of iterations, since the number of possible partitions of a finite set is finite, and at each step the total square deviation decreases, so looping is impossible.

In the proposed method, at each step, we select the largest cluster to maximize the f-beta-score metric as a harmonic mean between completeness and accuracy, which will allow us to obtain an average value between two significant metrics in the problem. Since there is no bijective mapping of clusters to previously obtained data in the K-means algorithm, the best set of labels of the target class is first assigned a label and then removed from the label pool. Not to lose information, we first consider the largest clusters, and then gradually move on to smaller clusters, since when a cluster is deleted, its label is lost.

Table 3. Label clustering matrix

	0	1	2	3	4	5	6
0	592	1,131	987	1,039	2,085	1,066	771
1	1	-	3,126	-	-	1	-
2	259	2,254	27	5	2	1	5
3	5	1	6,041	-	-	-	-
4	93	5	2,306	3	15	1	5
5	20	16	2,978	3	57	8	170
6	2	2	2,334	1	1	-	1

In our experiment we use a square partition matrix into seven clusters where we try to find the best non-parametric

partition for the data.

This approach is characterized by a strong overlap of classes. Therefore, when assigning the same target label to several classes at the same time, we use a greedy algorithm for selecting features, and we get a distribution of labels by classes, as in Table 3.

After calculating F_B (f-beta-score) using a formula of the harmonic mean:

$$F_B = (1 + B^2) * (PR * Recall) / (B^2 * (PR + Recall)) \quad (1)$$

where, PR is the accuracy, Recall is the completeness, and B is a configurable parameter of the offset between accuracy and completeness. We see that after a cycle of training and prediction, the f-beta-score exceeds the threshold of 0.52 only in classes I and IV.

This approach does not work well on the entire set of data, and changing the priorities of classes does not change anything since a significant number of points overlap each other.

When considering the clustering problem, we immediately note the impossibility of using the basic implementations of the density-based spatial clustering of applications with noise (DBSCAN) and ordering points to identify the clustering structure (OPTICS) algorithms [9]. We are interested in the full field of predictions, and these algorithms do not have a parameter for configuring the number of clusters because they are based on the ideas of availability points.

At this stage, we wanted to obtain $F_B \geq 0.91$. As we did not obtain it using embedding texts, we moved on to more advanced methods from the NLP field.

3.3 Recurrent approaches

Recurrent neural networks (RNNs) are very popular today in processing and analyzing data sequences, such as texts, time series, and audio or video streams. These architectures contain a cyclic layer that allows one to transfer information from the previous step in time to the next and to remember and use the previous contextual information for making decisions in the present. Formally, RNN is defined by the following equations:

$$h_t = \sigma h(W_x h * x_t + W_h h * h_{t-1} + b_h) \quad (2)$$

$$y_t = \sigma y(W_h y * h_t + b_y) \quad (3)$$

where, x_t is the input data at the current step, h_t is the hidden state at the current step, y_t is the output data at the current step, W and b are the matrices of weights and offsets, respectively, and σh and σy are activation functions for the hidden and output layers, respectively.

One of the NLP tasks, where RNNs have proven to be very effective, is to predict the diagnosis according to the description of thyroid disease. Using RNN, it is possible to analyze the results of the patient's studies and predict the diagnosis with high accuracy. A research team from the University of Leiden in the Netherlands conducted a study where they used an RNN to determine the diagnosis of thyroid diseases based on patient blood test data [9]. The study showed that RNN demonstrated higher prediction accuracy compared to traditional methods.

Last year, another group of researchers from Stanford University and Google Brain developed a new architecture for solving NLP problems, including predicting a diagnosis based

on thyroid analysis [10]. In this study, RNNs were used in combination with an attention mechanism that allowed the network to focus on the most significant elements of the sequence.

The results of the study also showed the high efficiency of the proposed method. The use of RNNs in the task of predicting the diagnosis based on thyroid analysis data allows us to obtain more accurate results and significantly reduce the diagnostic time, which is also confirmed by the author of the article on the use of neural networks in healthcare [10].

One of the most common approaches is the use of RNNs with LSTM cells. LSTM cells allow one to save information about previous network states and store it for a long time. An LSTM cell has the following components.

The use of LSTM cells allows an RNN to store and use information about previous states for a long time, which makes it especially effective for processing sequential data with long-term dependencies [11]. This is especially important when thyroid tests are collected several times over a certain period.

Another approach is the use of CNNs with 1D convolutions. This approach is especially useful when processing data that can be represented as time series, such as multiple thyroid assays over a certain period, which can be written as functions of time:

$$(exit) = Conv1D(input, filters) \rightarrow Activation \rightarrow Pooling \rightarrow Straightening \rightarrow Fully\ connected\ layers(output\ size)$$

In this case, 1D convolutions allow us to identify important temporal signs, such as peaks and trends, and use them to predict the diagnosis [12].

Finally, there are hybrid models combining several different neural network architectures that can be more efficient than each of them individually [13]. To illustrate this statement, a comparative analysis of the results of RNNs with LSTM cells, CNNs with 1D convolutions, and hybrid models combining both types of models was carried out. The results of the experiment, the purpose of which was to identify the most suitable solution for class label prediction are shown in Table 4.

Table 4. Model comparison

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
RNN + LSTM	80.7	81.5	79.9	80.7
CNN + 1D Conv	86.4	85.9	87.2	86.5
Hybrid Model	89.2	89.8	88.6	89.2

The table shows that the hybrid model, which combined RNN and CNN, showed the best results for all metrics. However, the CNN model with 1D convolutions came pretty close to the results and also showed great learning speed and ease of use.

3.4 Applying pre-trained models

Today, there are a large number of deep-learning models used for processing medical texts. The largest number of such models work in the English-speaking field of text analysis [14, 15]. However, some models have been pre-trained in Russian. To build such models, the well-known transformers BERT and

RoBERTa (Robustly Optimized BERT Approach) were used, which had already been pre-trained on common-language corpora [16].

In this paper, two other models are used, pre-trained on the Russian-language Wikipedia, and additionally on a special Taiga data corpus created by the SberDevice team [17, 18]. Both of these models were further transferably trained on publicly available medical and biomedical texts. Due to such training, RuBioBERT and RuBioRoBERTa models were obtained [19-21].

For a better understanding of the work of RuBioBERT and RuBioRoBERTa, we recall the basic principle of BERT. It consists of the fact that when training a neural network, a word is masked not only at the end of a sentence but also inside it. Such a task is called a masked language modeling task [14].

This approach allows the neural network to simultaneously learn (learn token embeddings) in both directions, thereby achieving deep bidirectionality, which means that the model considers the context on both sides of the word [22]. This architecture uses the attention mechanism, which helps to highlight the context and the relationship of words with each other.

The RoBERTa model is the same BERT model but with optimized hyperparameters and dynamic word masking. During the pre-training, RoBERTa is trained only to predict the masked word, whereas, in the BERT architecture, the model was also pre-trained on the task of next sentence prediction, i.e., it predicted whether the second sentence in a pair of sentences was a continuation of the first one [16, 23, 24].

For the task of determining the relationship between a cytopathological description and assigned Bethesda category, both of these models were used to understand which of them worked better when solving the classification problem. To compare RuBioBERT and RuBioRoBERTa, the following model parameters were used (Table 5).

Table 5. RuBioBERT and RuBioRoBERTa parameters

Parameter	Value
Learning rate	2e-5
Weight decay	0.01
Train epochs	5
Optimization	Adam

Table 6. The RuBioBERT learning process

Step	Train Loss	Validation Loss	Accuracy	F1
1,000	0.32	0.24	0.946	0.945
2,000	0.14	0.17	0.961	0.960
3,000	0.1	0.24	0.93	0.935
4,000	0.08	0.18	0.96	0.959
5,000	0.06	0.22	0.95	0.951

Table 7. The RuBioRoBERTa learning process

Step	Train Loss	Validation Loss	Accuracy	F1
1,000	0.26	0.20	0.953	0.952
2,000	0.14	0.19	0.961	0.961
3,000	0.10	0.17	0.970	0.969
4,000	0.08	0.20	0.960	0.960
5,000	0.06	0.19	0.953	0.961

The learning process in each of the models was recorded after every 1,000 steps and tested. As a result, the results of model training were obtained (Tables 6 and 7).

It can be seen from the tables that RuBioRoBERTa showed higher accuracy than RuBioBERT.

A significant difference between the RuBioRoBERTa model and RuBioBERT is a significant training time (3 times higher than the time of the other model) and a significantly greater weight of the model itself. Depending on the frequency of the necessary retraining, both the first and second models can be selected for classification according to Bethesda.

3.5 Text data augmentation

In the real corpus of 27,000 pairs of description and diagnostic category data used in this study, real descriptions are present only in 6,300 pairs. The description text can also be moved to the diagnostic category field. In the remaining data pairs, only a diagnostic category is present.

Considering that the quality of the model increases with an increase in the training sample, the task of augmenting the missing descriptions was additionally set. To do this, we trained a Russian-language T5-base model made by Sberbank to restore the text of a cytopathological description based on text of conclusion.

Augmentation (which took much more time than fine-tuning the T5-base model) increased the sample of pairs by almost 4 times to 24,600 examples. The analysis of the augmented data showed that:

- the average number of words in the original (36.2) and augmented (51.9) texts of descriptions did not differ much;
- the maximum number of words in the original texts was 177 and in the augmented texts 125.

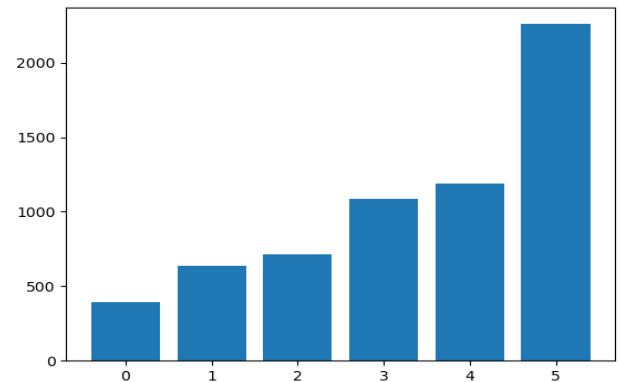


Figure 1. Distribution of descriptions within topics in real descriptions

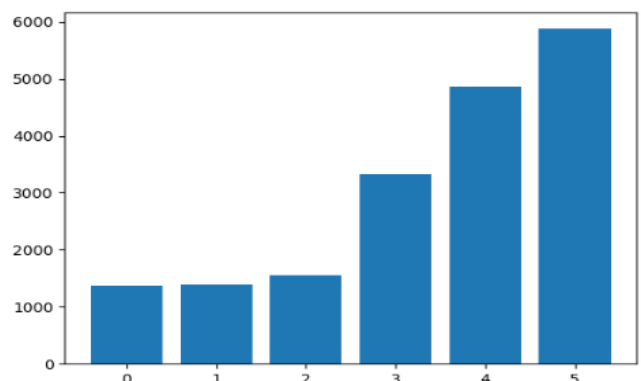


Figure 2. Distribution of descriptions within topics in augmented descriptions

To analyze the properties of the source and synthetic data, thematic modeling of LDA (Latent Dirichlet Placement) was carried out using the genism library, where the initial class (the first category according to Bethesda) is marked with zero and further increases. The number of topics was taken equal to 6, according to the number of Bethesda categories.

The distribution was sorted by increasing probabilities to be able to see the general pattern. The distributions of the original and synthetic data are very similar (Figures 1 and 2).

In addition to the distribution of the descriptions by topic, the distribution of topics by words was constructed to highlight the top words describing a specific topic (Tables 8 and 9).

Table 8. Top tokens of the original descriptions

Topic	Top Words
0	Cell, detect, gland, thyroid, background
1	Goiter, colloidal, nodular, punctured, element
2	Cell, cluster, background, preparation, position
3	Quantity, small, smear, thyrocyte, background
4	Epithelium, cell, follicular, detect, background
5	Allow, recommend, observation, dynamic, nodal

Table 9. Top tokens of synthetic descriptions

Topic	Top Words
0	Element, blood, peripheral, naked, nucleus
1	Erythrocyte, rounded, nucleus, monomorphic, relatively
2	Element, goiter, colloidal, nodular, lymphocytic
3	Characteristic, cytogram, goiter, colloidal, change, follicular
4	Papillary, thyroid, gland, erythrocyte
5	Erythrocyte, change, background, group, colloid

The vocabulary in the topics almost does not match, but structurally the descriptions are similar.

4. RESULTS

Fine-tuning of the Russian-language T5-base model to solve the problem of generating a diagnostic category based on a description was carried out on augmented (24,600 examples) and initial (6,200 examples) samples.

Then, with the help of a trained model, predicted diagnostic categories were generated based on the available source descriptions, using regular expressions. According to the initial and predicted cytopathological descriptions, the Bethesda classes were obtained.

The accuracy of training and validation of the neural network model used for automatic classification of the disease description by Bethesda is shown in Figure 3.

To assess the quality of the created models of automatic generation of diagnostic categories on the text of the FNA description, a set of metrics for summarizing the text of ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (ROUGE-1, ROUGE-2, ROUGE-L) was used.

It is considered that the indicators ROUGE-2 and ROUGE-L are well suited for the tasks of abstracting individual documents, and ROUGE-1 and ROUGE-L show good results in evaluating short abstracts. Considering that the medical report can be both short and medium in length, we used all three indicators.

ROUGE automatically generated descriptions with references or a set of references (human created descriptions).

ROUGE scores range from 0 to 1, with higher scores indicating greater similarity between automatically generated descriptions and references. Examples of formulas for calculating ROUGE:

$$ROUGE - 1 = \frac{Num\ word\ matches}{Num\ word\ in\ reference}$$

$$ROUGE - 2 = \frac{Num\ bigram\ matches}{Num\ bigrams\ in\ reference}$$

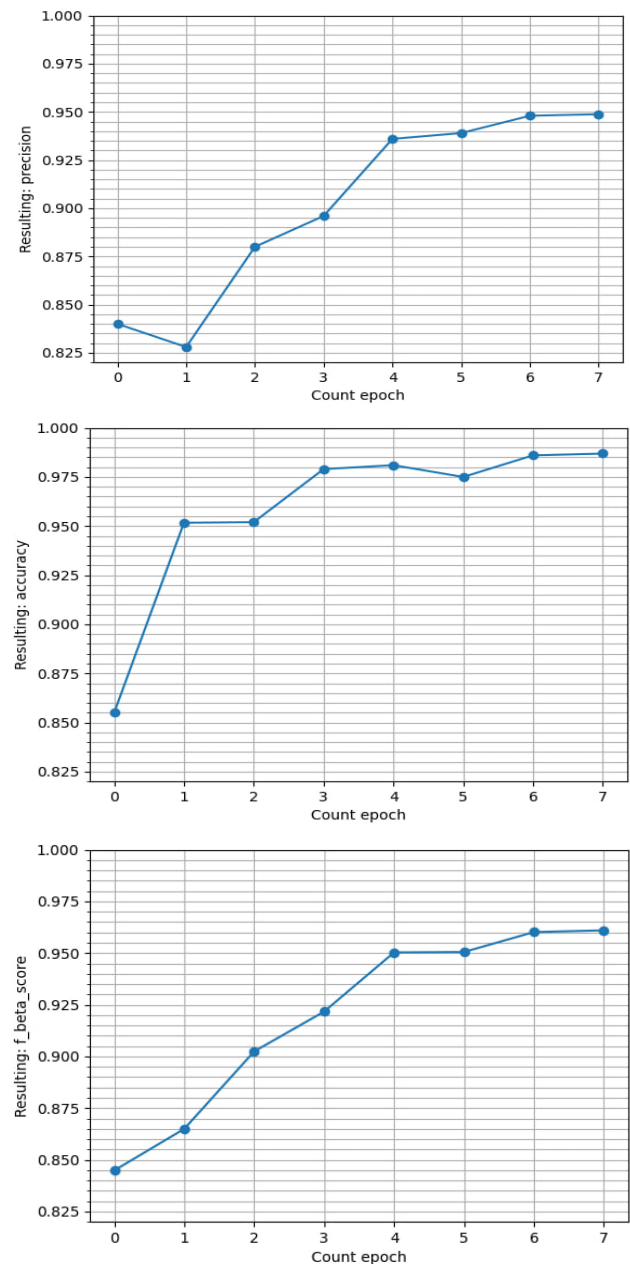


Figure 3. Graphs of the accuracy of training and validation of the neural network model in classification

The key idea of the ROUGE family metrics is to estimate the intersection by n-grams between the source text and the generated one (in our case, between the text written by the doctor and the text generated by the model). The ROUGE-1 metric considers only individual words (unigrams), adhering to the bag of words philosophy, which does not allow evaluating the quality of word combinations.

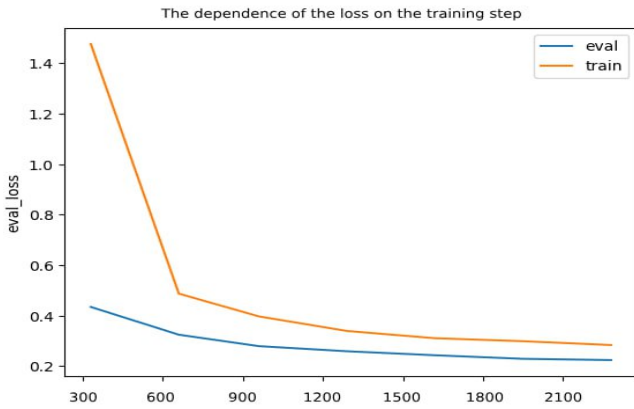


Figure 4. Graph of the loss function

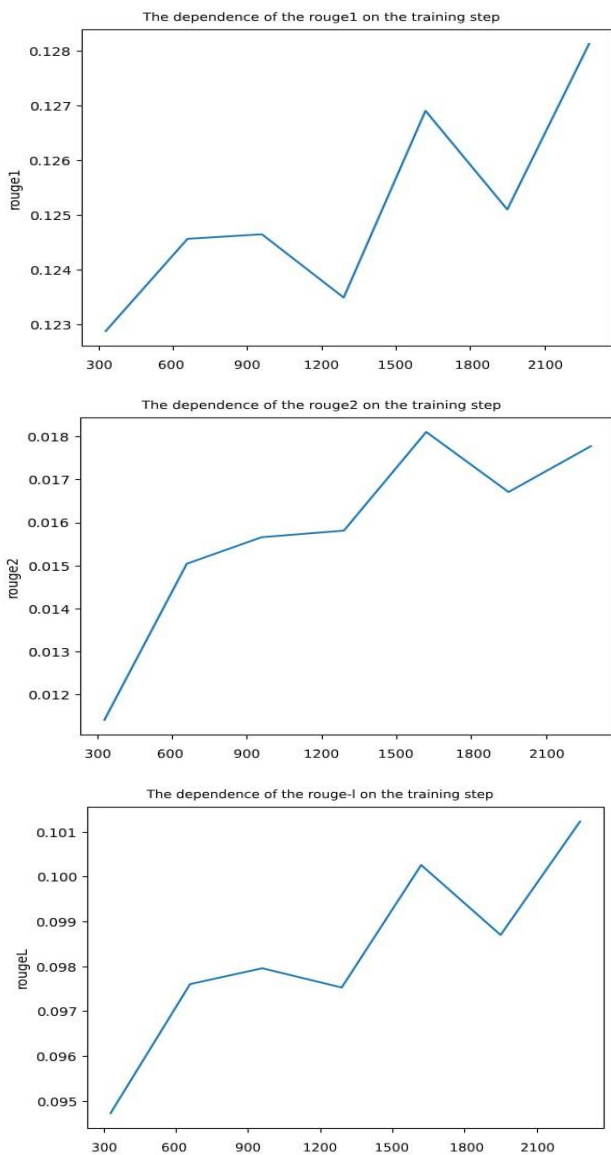


Figure 5. Graphs of ROUGE summarization metrics

This problem is partially eliminated by ROUGE-2 and ROUGE-L, as they evaluate matches by bigrams and the longest sequence of words, respectively. The last two metrics are more demanding to text generation than ROUGE-1. The change in the logistic loss function (loss) and the metrics of the ROUGE family depending on the step of the batch

optimization of the neural network architecture can be seen in Figures 4 and 5.

The texts of most of the generated medical assessment reports coincide with the texts of the medical assessment reports written by doctors. Examples of differing medical assessment reports are given in Table 10.

Table 10. Examples of medical assessment reports

Suggested by the Doctor	Generated
L p/p 21. A benign nodule of the thyroid gland, diagnostic category II according to the criteria of the Bethesda classification system. L n/3 21. An atypical structure of unclear significance, diagnostic category III according to the criteria of the Bethesda classification system	L n/3 21. An atypical structure of unclear significance, diagnostic category III according to the criteria of the Bethesda classification system. L n/3 21. Benign nodules of both lobes of the thyroid gland, diagnostic category II according to the criteria of the Bethesda classification system
Cells of a neoplasm suspected of malignancy (papillary thyroid carcinoma?)	Suspicion of a malignant neoplasm, diagnostic category VI according to the criteria of the Bethesda classification system
A neoplasm suspected of malignancy, diagnostic category V according to the criteria of the Bethesda classification system	A thyroid neoplasm of unclear malignant potential, diagnostic category IV according to the criteria of the Bethesda classification system

Now we can conduct statistical tests for the two subproblems being solved, classification and generation. In the first case, the value will be f-score, in the second the resulting Rouge-2 values, the values are presented in Table 11.

Table 11. TestInd

Type	F1 Identic	ROUGE2 Identic	P Value
Classification	0.09	-	0.05
Generation	-	0.05	0.08

For classification values, we took the basic p-value threshold for text generation tasks, we set more stringent measures and take the threshold for accepting the null hypothesis equal to 0.08.

5. DISCUSSION

Methods of converting natural language texts for medical purposes are currently being actively developed.

The tasks being solved are difficult in terms of data processing and further training of the model since errors in the domains of medical tasks are critical. This forces us to collect data more carefully with an understanding of what results we want to achieve with augmentation using ready-made or trained models. Approaches to generating and evaluating data arrays are discussed in papers [9, 10].

However, in our task, based on a corpus of ready-made texts, the main focus was on converting raw data and not on attempts to anonymize the personal data of patients, which had been done by other people in advance. The ideas of preprocessing formed the basis for obtaining data clusters but this did not show the expected tangible increase in target metrics in the future.

In the broader context of text classification, similar studies have encountered challenges related to preprocessing efficacy. For instance, the studies [25, 26] faced a comparable situation where traditional preprocessing methods alone did not yield significant advancements in target metrics. This parallel underscores the inherent intricacies of text data, particularly in medical narratives, and the need for nuanced preprocessing strategies.

The main approaches that could help solve the problem were based on the sequential generation of matrix kernels in recurrent networks and architectures based on convolutions. These approaches are not new but necessary for generating the required labels of the class under study, and the ideas of the applied methods are partially covered in study [13, 14].

The approaches proposed in the paper to solving the problems of data preprocessing, classification, and augmentation proved to be good at generating Bethesda labels but were not effective enough for complex token sequences.

This leads to an understanding of the issue of insufficient data in the training sample. To solve it, attempts were made to use ready-made pre-trained architectures with additional training on the target data set. Such techniques are based on the algorithms proposed in study [16].

Comparing our approach with similar studies in the field, we observed that [27] encountered a similar data insufficiency issue in their research. They addressed this challenge by employing a transfer learning strategy, achieving significant improvements in their model's performance. In our case, leveraging pre-trained architectures allowed us to navigate the data limitations, resulting in enhanced accuracy and robustness in handling complex token sequences.

To obtain the text of a complete description, as a result of generating a sequence of tokens of a medical opinion, previously pre-trained BERT models can be used. However, to expand the feature space, data augmentation and additional training of transformer models on the existing corpus of texts are required. This is what will allow for obtaining more accurate and complete medical reports at the output of the model.

In our study, we employed two primary models for the classification of thyroid cytopathology reports in the Russian language: Sequential neural networks and Transformer neural networks. Both models demonstrated commendable performance, but a closer examination reveals nuanced differences in their effectiveness.

The choice of model depended on the scale of the dataset. Sequential neural networks shone in resource-constrained scenarios, offering robust performance, while Transformer neural networks demonstrated their prowess when dealing with larger datasets, where their ability to handle complex token sequences yielded even higher accuracy.

6. CONCLUSIONS

This study addressed the challenge of processing Russian-language thyroid cytopathology reports using deep learning models. We utilized Sequential and Transformer neural networks, comparing their effectiveness in disease classification and report generation across different dataset sizes.

Our findings revealed that Sequential neural networks are more suited for smaller datasets, showing high accuracy and efficiency. In contrast, Transformer models demonstrated

superior performance with larger datasets, excelling in complex text processing tasks. These insights are crucial for advancing the application of deep learning in medical text analysis, particularly in non-English contexts.

The result of this work was the creation of a pipeline. At its first stage, statistical methods were used to clean and preprocess raw data, which worked faster than neural networks with almost the same qualitative result.

Next, models of a transformer operating in two modes were trained as a Bethesda label generator and as a medical assessment report token generator based on the input description.

The proposed pipeline has an interface for preprocessing and retraining the model and can be converted into a serializable .pkl file.

Since the work was done in a high-level Python language, the final developments can be scaled. With the help of sklearn libraries to scale the interface and with the help of transformers to replace individual steps of the converter of the tasks being solved, the methods of preprocessing data from other medical fields can be left unchanged.

In terms of research metrics, an improvement of 2% was achieved, which is due to the proposed approaches to processing Russian-language data and combining methods of retraining ready-made architectures.

The practical significance of our results lies in their potential to make the classification of thyroid cytopathology reports easier, reduce manual labor, optimize resources, and pave the way for improved healthcare efficiency.

This work can be implemented in systems that rely on the automation of medical predictions and can help doctors make predictions much faster than using manual methods.

Considering the ethical component of this work, we note that the resulting pipeline is not yet a replacement for a doctor but is positioned as an assistant and interlocutor when forming a medical assessment report on the patient's problems.

ACKNOWLEDGMENTS

The authors would like to thank the Institute of Physics and Biology of the National Research Nuclear University MEPhI for support of this study.

REFERENCES

- [1] Ali, S., Cibas, E. (2018). The Bethesda system for reporting thyroid cytopathology. Springer International Publishing, Cham.
- [2] Ali, S.Z., Baloch, Z.W., Cochand-Priollet, B., Schmitt, F.C., Vielh, P., VanderLaan, P.A. (2023). The 2023 Bethesda System for reporting thyroid cytopathology. *Thyroid*, 33(9): 1039-1044. <https://doi.org/10.1089/thy.2023.0141>
- [3] Wang, Z., Ive, J., Velupillai, S., Specia, L. (2019). Is artificial data useful for biomedical Natural Language Processing algorithms. In: Proceedings of the 18th BioNLP Workshop and Shared Task. Association for Computational Linguistics, Florence, pp. 240-249. <https://doi.org/10.18653/v1/W19-5026>
- [4] Vasilyev, O., Bohannon, J. (2022). Neural embeddings for text. arXiv preprint arXiv:2208.08386. <https://doi.org/10.48550/arXiv.2208.08386>

- [5] Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H. (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1): 2. <https://doi.org/10.1145/3458754>
- [6] Houssein, E.H., Mohamed, R.E., Ali, A.A. (2021). Machine learning techniques for biomedical natural language processing: A comprehensive review. *IEEE Access*, 9: 140628-140653. <https://doi.org/10.1109/ACCESS.2021.3119621>
- [7] Miolo, G., Mantoan, G., Orsenigo, C. (2021). Electrmed: a new pre-trained language representation model for biomedical nlp. *arXiv preprint arXiv:2104.09585*. <https://doi.org/10.48550/arXiv.2104.09585>
- [8] Naseem, U., Khushi, M., Reddy, V., Rajendran, S., Razzak, I., Kim, J. (2021). Bioalbert: A simple and effective pre-trained language model for biomedical named entity recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, Shenzhen, China, pp. 1-7. <https://doi.org/10.21203/rs.3.rs-90025/v1>
- [9] Mody, D.R., Thrall, M.J., Krishnamurthy, S. (2018). *Diagnostic Pathology: Cytopathology* (2nd ed). Elsevier.
- [10] Lyu, C., Chen, B., Ren, Y., Ji, D. (2017). Long short-term memory RNN for biomedical named entity recognition. *BMC Bioinformatics*, 18(1): 462. <https://doi.org/10.1186/s12859-017-1868-5>
- [11] Fartushny, E.N., Sych, Yu.P., Fartushny, I.E., Koshechkin, K.A., Lebedev, G.S. (2022). Stratification of thyroid nodes by eu-tirads categories using transfer learning of convolutional neural networks. *Clinical and Experimental Thyroidology*, 18(2): 17-26. <https://doi.org/10.14341/ket12724>
- [12] Van Houdt, G., Mosquera, C., Nápoles, G. (2020). A review on the long short-term memory model. *Artificial Intelligence Review*, 53: 5929–5955. <https://doi.org/10.1007/s10462-020-09838-1>
- [13] Alzubaidi, L., Zhang, J., Humaidi, A.J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaria, J., Fadhel, M.A., Al-Amidie, M., Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8(1): 53. <https://doi.org/10.1186/s40537-021-00444-8>
- [14] Vyucheyanskaya, M.V., Krainova, I.N., Gribanov, A.V. (2018). Neural network technologies in the diagnosis of diseases (review). *Journal of Biomedical Research*, 3: 284-294. <https://doi.org/10.17238/issn2542-1298.2018.6.3.284>
- [15] Blinov, P., Reshetnikova, A., Nesterov, A., Zubkova, G., Kokh, V. (2022). RuMedBench: A Russian medical language understanding benchmark. In: Michalowski, M., Abidi, S.S.R., Abidi, S. (eds) *Artificial Intelligence in Medicine. AIME 2022. Lecture Notes in Computer Science*, vol. 13263. Springer, Cham, pp. 383-392. https://doi.org/10.1007/978-3-031-09342-5_38
- [16] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In: Burges, C.J., Bottou, L., Welling, M., Ghahramani, Z. and Weinberger, K.Q. (eds) *Advances in Neural Information Processing Systems*. Curran Associates, Inc., Red Hook, 3111-3119.
- [17] Pennington, J., Socher, R., Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, pp. 1532-1543.
- [18] Devlin, J., Chang, M., Lee, K., Toutanova, K. (2019). BERT: Pretraining of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Association for Computational Linguistics, Minneapolis, pp. 4171-4186. <https://doi.org/10.18653/v1/N19-1423>
- [19] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. (2018). Improving language understanding with unsupervised learning. Technical report, OpenAI. <https://openai.com/research/language-unsupervised>
- [20] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I. (2017). Attention is all you need. In: *Advances in Neural Information Processing Systems*. Curran Associates, Montreal, pp. 6000–6010.
- [21] Peng, Y., Yan, S., Lu, Z. (2019). Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking In: *Proceedings of the 18th BioNLP Workshop and Shared Task*. Association for Computational Linguistics, Florence, Italy, pp. 58-65. <http://dx.doi.org/10.18653/v1/W19-5006>
- [22] Akhmetshin, E., Nemtsev, A., Shichiyakh, R., Shakhov, D., Dedkova, I. (2024). Evolutionary algorithm with deep learning based fall detection on internet of things environment. *Fusion: Practice and Applications*, 14(2): 132-145.
- [23] Yalunin, A., Nesterov, A., Umerenkov, D. (2022). RuBioRoBERTa: a pre-trained biomedical language model for Russian language biomedical text mining. *arXiv preprint arXiv:2204.03951*. <https://doi.org/10.48550/arXiv.2204.03951>
- [24] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J. (2019). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234-1240. <https://doi.org/10.1093/bioinformatics/btz682>
- [25] Lawson, C.E., Martí, J.M., Radivojevic, T., Jonnalagadda, S.V.R., Gentz, R., Hillson, N.J., Martin, H.G. (2021). Machine learning for metabolic engineering: A review. *Metabolic Engineering*, 63: 34-60. <https://doi.org/10.1016/j.ymben.2020.10.005>
- [26] Vora, L.K., Gholap, A.D., Jetha, K., Thakur, R.R.S., Solanki, H.K., Chavda, V.P. (2023). Artificial intelligence in pharmaceutical technology and drug delivery design. *Pharmaceutics*, 15(7): 1916. <https://doi.org/10.3390/pharmaceutics15071916>
- [27] Zhao, Z., Alzubaidi, L., Zhang, J., Duan, Y., Gu, Y. (2024). A comparison review of transfer learning and self-supervised learning: Definitions, applications, advantages and limitations. *Expert Systems with Applications*, 242: 122807. <https://doi.org/10.1016/j.eswa.2023.122807>