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

Unstructured Text and Tabular Information Processing in the Clinical Decision Making System for the Respiratory Diseases Diagnosis

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

A.A. Evgrafov*
Computer Science and Robotics Department
Ufs State Aviation Technical University
Ufa, Russia
evgrafov.alexander92@yandex.ru

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

***Abstract*—The article considers the basic algorithm of text and table data extraction used in the developed system of clinical decision-making in diagnosis of respiratory diseases, methods of formation of the data structure of an individual patient, a set of data from all patients for further application in models of machine learning as well as construction of ML models which provide detection of disease in patients. Data extraction and generation processes are performed in the Python programming language using additional libraries: "docx" and "pandas" for data processing and "sklearn", "lightgbm" and "catboost" for building machine learning models. The relevance of the task is due to large volumes of unstructured data received by the CDSS input and necessary for its effective functioning. The novelty of development lies in application of a set of existing and development of new algorithms of extraction and primary processing of text and table information.**

Keywords—clinical decision support system, data processing, processing of text information, text mining, machine learning

I. INTRODUCTION

The active penetration of information technologies into medicine and health care has stimulated development of software complexes with the main purpose of improving the quality of medical services by introducing modern intelligent technologies and machine learning technologies into the process of interaction of medical personnel with patients. One of the most effective tools for achieving this goal is the use of clinical decision-making systems (CDSS). CDSS may be defines as systems designed to help physicians and other medical professionals work with clinical decision-making tasks focus on processing and managing information to make recommendations in patient treatment based on available data [1]. Application of CDSS assures higher efficiency of diagnosis and treatment of the patient as well as reduces the burden on the doctor during information processing.

The [2] describes how the architecture of the system for making clinical decisions in the diagnosis of respiratory diseases was developed. This system involves such functions as analysis of patient data in the form of text, graphic and sound in-formation, presentation of the expected diagnosis on the basis of the analysis carried out and the detected patterns, formation and issuance of medical recommendations to increase the patient's treatment efficiency [3].

The purpose of this article is to consider, describe and implement the basic algorithm required for:

- uploading patient data as docx files;
- forming a single data structure for each individual patient;
- filling in the created data structure for each patient;
- generating a unified data set for all patients;
- using machine learning algorithms to predict the presence of disease in the patient.

The implemented algorithm is a component of the text data recognition unit of the developed clinical decision support system.

II. STATE OF ART

Today, many technology companies are developing systems related in one way or another to medicine and unstructured text processing.

For example, Amazon is expanding its Comprehend developer service for natural language processing (NLP), which uses machine learning technologies to find patterns and relationships in text. The product, called Amazon Comprehend Medical, will extract and structure meaningful information (complaints, diagnosis, prescribed drugs and their dosage, research results, etc.) from unstructured medical records. This does not require machine learning knowledge from the user, as Comprehend Medical is provided by the service model through Amazon Integration Services (APIs) [4].

Developers from Google solve the task of depersonizing audio data using the algorithm of automatic determination of named entities (NER - Named-Entities Recognition), which is used for search and classification of proper names, names of events, products, toponyms, etc. Among them can be names of people or companies, geographical objects (cities, streets, etc.). A feature of this system is that types are not defined by a dictionary, but assigned on the basis of statistical algorithms. The process of deleting personal data

starts with automatic speech recognition and translation into text (ASR - Automatic Speech Recognition), then NER marks all data that may relate to personal data, and they are deleted from the text, after which the text is translated again into audio recording [5].

Another development of the recent period was the database of the National Medical Library (NLM), which uses artificial intelligence technologies to provide prompt reporting and issuance of up-to-date literature and COVID-19 resources for researchers and scientists. As of May 1, 2020, the developed model of the library included about 46 thousand articles obtained from PubMed Central (PMC) and ClinicalTrials.gov. Development uses machine learning for textual information and is likely to accelerate COVID-19 research [6].

It should be noted that, if necessary, open-source modules for processing large volumes of textual information created by third-party companies can be considered and used in the developed CDSS, but for this purpose these modules are subject to mandatory adjustment and refinement due to the specifics of processing medical data that do not have a single structure and are oriented to documents written in Russian with the presence of Latin terminology, for which the services offered are not initially oriented and the result is not obvious. What follows is the need to develop their own algorithms and modules that take into account the specifics of the CDSS data.

III. Statement of the Problem

The main problem of text information processing in the developed CDSS is its unstructured form. The data are presented in the form of sentences in natural languages as well as tables containing numerical and categorical values.

Processing unstructured text requires the use of sufficiently sophisticated auxiliary means to extract useful information, such as lemmatisers, linguistic analyzers, regular expressions, and, in some cases, pre-learned models of neural networks, such as BERT.

A further challenge in the development of the CDSS in the diagnosis of respiratory diseases is the need to create a data structure that meets the requirements of fast and easy access. The system being developed involves processing a large array of information about various patients, so that a well-formed structure will simplify the development of CDSS and reduce the requirements for resources consumed during its operation.

Currently, textual information enters the clinical decision-making system in the form of text documents with extension * .docx and is presented by statements from patient medical histories. A fragment of one patient's statement is shown in Figure 1.

The data used in the kit was previously anonymized and does not contain any personal information not relevant to the state of patient‘s health. The dataset includes typical patient information:

- full name of the patient (changed);
- dates of birth, admission and discharge from the hospital;
- complete diagnosis, complication information;
- tabular data with diagnostic tests;
- text data with diagnostic studies expressed in an unstructured form;
- description of the treatment methodology and applied drugs;
- recovery recommendations;
- hospitalization result.

When developing the algorithm, special attention is paid to the formation of the data structure pertaining to the future dataset for machine learning models and the preprocessing of tabular information with diagnostic studies. Text information is preliminarily placed in a separate line of the structure for the possibility of working with it in the future.

Table information is represented by five types of tables:

1. General blood test (CBC).
2. General urine analysis (CUT).
3. Blood chemistry (BBA).
4. Coagulogram (BCT).
5. Sputum. General analysis (ACE).

For individual patients, the variability of the tables may vary slightly.

IV. Proposed Decision

According to the available data, a basic text data structure has been developed in CDSS (Figure 2), which will structure information about each patient and provide easy and quick access to the rest of the data-oriented modules of the system.

The basic structure of text data contains a list of patients, inside which, in turn, dictionaries are placed, separating text

```
extract from the medical history
surname, name, patronymic: ivanova veronika andreevna
dob:                              12.08.1995 г.р.
date of admission to the hospital:      20.09. 2019 г.
date of discharge from hospital:        30.09. 2019 г.
full diagnosis (underlying disease, complication, concomitant):
primary: community-acquired pneumonia with localization in the upper lobe of the left l
ung, moderate severity.
complication: respiratory failure of 1 degree.
```

Fig. 1. A fragment of a patient history statement

and table structures of data. In the future, the number of dictionaries can be increased by additional information about the patient, such as images (CT and MRI images) or audio information (auscultation results). The content of a dictionary containing text information is data in string format.

Text information is extracted according to the following algorithm: for each line of the source document concatenation is performed with pre-bringing the line to the

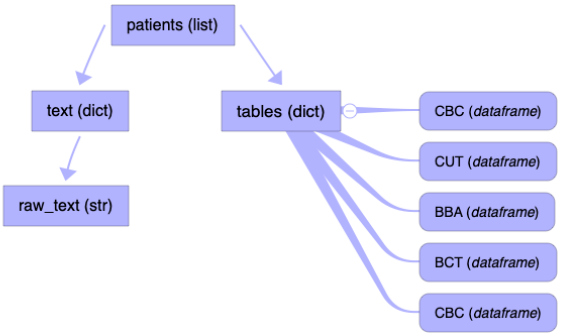

Fig. 2. The basic structure of text data developed by CDSS

lower case, which will allow to exclude influence of upper case characters on further text processing.

To store tables, use the *dataframe* type represented by the software library for processing and analyzing *pandas* data [7]. *Pandas* assumes a fairly high speed and ease of manipulation of indexed arrays of two-dimensional data. In case of substantial expansion of patient dataset, transition from pandas to low-level, and therefore faster library for working with arrays - *numpy* [8] is assumed. Table-like data extraction and data set generation takes place according to the following algorithm:

1. Each of the tables is composed of separate text cells.
2. The first cell of the table defines its name and places the table in a dataframe containing the names of the columns corresponding to the table name.
3. Cells with no values are converted to "NaN".
4. Cells containing dates are converted to datetime.
5. For values in the numeric columns, symbols «,» must first be replaced by «.» because of the difference in the recording standards, and then converted to float format.

The result of the algorithm in the form of a table of general blood analysis of one of the patients is shown in Figure 3.

Model tables are created for each patient, but the application of machine learning technologies is based on the sets of all available data, so the next step of information processing is to combine the previously formed data.

Since the CDCC under development involves the possibility of diagnosis, the primary task is to divide patients into two classes: sick and healthy. The solution of this task will allow to create more accurate models for diagnosis of the sick in the future.

The table in Figure 3 contains two lines, the first of which includes the results of patient's tests when entering the hospital (i.e., the sick person), and the second one shows results when leaving the hospital, which should correspond to the state of a healthy person. By grouping and aggregating the data, we will distribute the results of patient tests to the healthy and sick. We will also remove from the list the characteristics with a considerable amount of gaps.

We will perform similar operations for the table containing the general urine analysis. For the rest of the tables, conversion of the data will not be possible due to the one-time analysis of the patient (upon admission to hospital), so the use of tables containing data from biochemical blood analysis, coagulogram, and general sputum analysis will be used to further clarify the most likely diagnosis of the patient.

In order to form the final set of data and use it in the machine learning model, we will combine the previously formed tables of general blood test and general urine test (the result is shown in Figure 4).

The target feature in the model for classifying patients into sick and healthy patients is the "sick" parameter. As characteristics for model training and testing, it is decided not to use categorical characteristics such as "epith," “pro”, "col," “transp”, "wbc_y", as in the future they will require additional processing. The following are used as machine learning models in the primary stage:

- Logistic Regression [9];
- Decision Tree Classifier [10];
- Random Forest Classifier [11];
- LightGBM Classifier [12];
- CatBoost Classifier [13].

The test sample size is 30%. The results are obtained by cross-validation with the number of data splits into 5 samples. The best quality is achieved using the following list of characteristics:

- «segm» – relative number of segmentonuclear neutrophils;
- «wbc_x» – leukocytes;
- «ph»-reaction / acidity;
- «sg» - specific weight.

Metrics used are:

- accuracy – percentage of values for which the classifier has made the correct decision;

	cbc_date	rbc	hb	wbc	mchc	esr	eos	stab	segm	lym	mon	bas	rpr	young
0	2019-09-24	3.97	119.0	8.2	NaN	19.0	1.0	2.0	59.0	34.0	4.0	NaN	NaN	NaN
1	2019-09-27	4.30	131.0	11.7	NaN	9.0	1.0	3.0	77.0	16.0	3.0	NaN	NaN	NaN

Fig. 3. Example of a table for general blood analysis of one of the patients

- precision – the proportion of objects actually belonging to a given class relative to all the values that the system has assigned to that class (the proportion of really sick patients among all patients labeled as "sick");

leading to the need for further adjustment and improvement of the model.

With a significant increase in learning sampling, consideration should be given to moving from a random forest model to models using gradient booster (LightGBM

	patient	cbc_date	esr	hb	lym	mon	rbc	segm	stab	wbc_x	sick	col	cut_date	epith	ph	pro	sg	transp	wbc_y
9	13	2019-02-07 00:00:00	21.0	146.0	28.0	7.0	5.20	62.0	3.0	5.7	1	с/ж	2019-02-07 00:00:00	един	5.0	отриц	1025.0	прозр	1-0-1
4	4	2018-10-18 00:00:00	38.0	148.0	11.0	2.0	4.75	84.0	3.0	12.2	1	с/ж	2019-10-18 00:00:00	NaN	5.0	2.1	1036.0	прозр	2-4-5
18	3	2019-01-22 00:00:00	8.0	135.0	42.0	7.0	4.84	45.0	2.0	5.3	0	с/ж	2019-01-22 00:00:00	2-3-3	5.5	отриц	1018.0	прозр	един
1	1	2019-09-02 00:00:00	45.0	132.0	29.0	4.0	3.80	64.0	2.0	5.8	1	с/ж	2019-09-02 00:00:00	един	5.0	отриц	1026.0	прозр	NaN
0	0	2019-09-24 00:00:00	19.0	119.0	34.0	4.0	3.97	59.0	2.0	8.2	1	с/ж	2019-09-23 00:00:00	1-1-2	5.0	отриц	1015.0	прозр	един

Fig. 4. A combined table containing data from all patients

- recall – shows what proportion of objects really belonging to the positive class is predicted to be true (the proportion of really sick patients who have been defined as "sick" among all sick patients);
- F-measure – average harmonic value of accuracy and completeness [14].

The obtained results calculated on training and test samples are shown in Figures 5 and 6. The best indicators,

and CatBoost) because the quality of these models is generally significantly improved using a large number of data. With a significant increase in the number of characteristics, the characteristics that most affect the quality of the model can be highlighted using the Feature Import mechanism, or the Shap library [15], and minor characteristics are removed.

Also one of the forthcoming tasks to improve the quality of the model is to process categorical characteristics and

	model name	accuracy train	precision_train	recall train	f1 train	accuracy test	precision_test	recall test	f1 test
0	Logistic Regression	0.74	0.77	0.80	0.76	0.67	0.50	1.00	0.67
1	Decision Tree Classifier	0.69	0.77	0.70	0.69	0.78	0.67	0.67	0.67
2	Random Forest Classifier	0.77	0.78	0.83	0.79	0.78	0.60	1.00	0.75
3	LightGBM Classifier	0.57	0.57	1.00	0.73	0.33	0.33	1.00	0.50
4	CatBoost Classifier	0.70	0.77	0.80	0.73	0.78	0.67	0.67	0.67

Fig. 5. Evaluation of models on the training and test sample

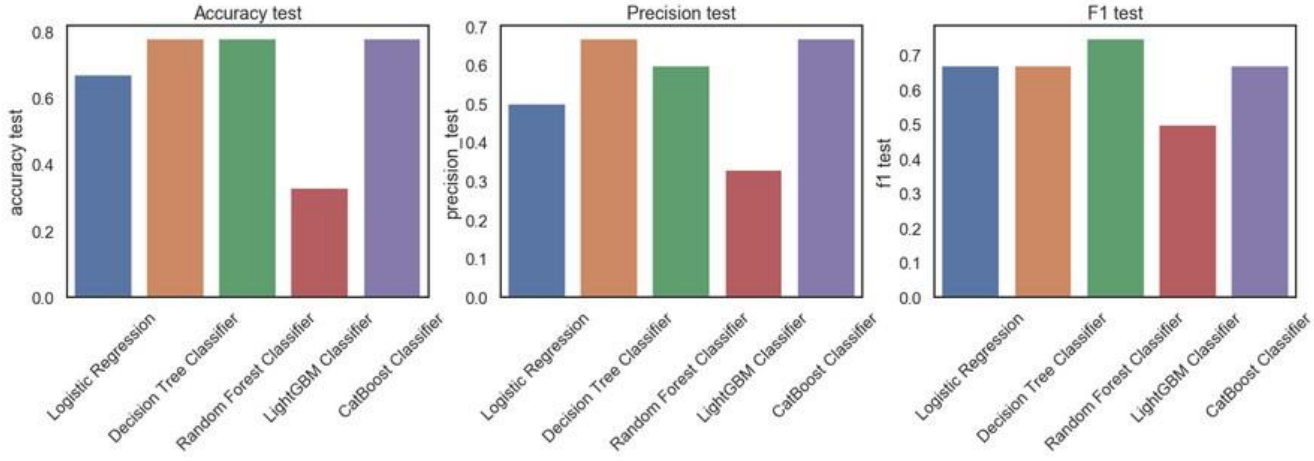

Fig. 6. Comparison of accuracy, precision, F1 metrics for different models on the test sample

according to the F1 metric, are shown by the Random Forest classifier. A recall value of 1 on the test sample suggests no false-negative results, meaning none of the really sick patients are defined as healthy. However, since the precision of the model is 0.6, this means that a substantial number of healthy patients from the sample are also labeled as sick,

compile dictionaries of possible values for them. Characteristics of this type are processed by means such as LabelEncoder, OneHotEncoder. Processing of categorical variables can be carried out, among other things, by built-in CatBoost library tools.

More detailed decryption of tabular data and implementation of developed processing algorithm are located in Github repository [16].

V. CONCLUSIONS

The article discusses the algorithm of processing unstructured textual and tabular information obtained from the history of diseases, developed for the system of supporting clinical decision-making in order to create a classifier, which separates patients into sick and healthy:

- A basic structure of text and tabular data has been developed to ensure easy storage and quick access to patient information.
- An algorithm has been developed to extract information from the medical history statements of individual patients contained in separate files of the * docx format, which allows to use without additional transformations those patient data that are currently stored in most Russian medical institutions.
- A part of the extracted data is combined and a single dataset is created to prepare it as input information for machine learning models.
- The models of classifiers based on regression, decision trees and gradient booster were evaluated, and, subsequently, the best result from the current data was shown by ensembles of decision trees. The best classifier has been chosen, allowing to monitor the condition of patients upon receiving their test results and recording the moment of recovery. Special attention is paid to the processing of tabular information, including the results of patient tests. Based on the processed data, machine learning models have been built, the best model recognized is the Random Forest classifier, which reached 0.75 F1-metrics on the test sample.
- Directions for further improvement of classifier quality are identified.

The developed models and algorithms will allow to classify the presence or absence of disease on the basis of the initial analysis of patient data and further, on the basis of the developed structure, taking into account the data extracted from unstructured information, to move to the development of modules of in-depth analysis of data whose functionality will provide an opportunity to establish the most probable diagnosis for each patient.

ACKNOWLEDGMENT

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

REFERENCES

[1] "Medical Decision Support System" [Electronic Resource], URL: http://www.unix-medical.ru/Katalog-produkcii/IT-reshenija-v-medicine/Iskusstvennij-intellekt-v-medicine/46913.html (date of the request: 25.05.2020).

[2] Shakhmametova G.R., Clinical Decision Support System for the Respiratory Diseases Diagnosis / Shakhmametova G.R., Zulkarneev K. Kh, Evgrafov A.A; Proceedings of the 7th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2019), Atlantis-Press [Electronic Resource], URL: https://www.atlantis-press.com/proceedings/itids-19/125908965 (date of the request: 31.05.2020).

[3] Shakhmametova G.R., Analytical Processing of Medical Data in the System of Diagnosis of Bronchopoietic Diseases/Shakhmametova G.R., Zulkarneev R.H., Evgrafov A.A.; ITIDS'2018, Volume 1, c.256-261 (2018).

[4] AWS Machine Learning Blog - Introducing medical language processing with Amazon Comprehend Medical, [Electronic Resource], URL: https://aws.amazon.com/ru/blogs/machine-learning/introducing-medical-language-processing-with-amazon-comprehend-medical/ (date of the request: 02.06.2020).

[5] Ido Cohn, Itay Laish, Genady Beryozkin - Audio De-identification: A New Entity Recognition Task, [Electronic Resource], URL: https://arxiv.org/abs/1903.07037 (date of the request: 02.06.2020).

[6] Melissa Harris - NLM Leverages Data, Text Mining to Sharpen COVID-19 Research Databases, [Electronic Resource], URL: https://governmentciomedia.com/nlm-leverages-data-text-mining-sharpen-covid-19-research-databases (date of the request: 04.05.2020).

[7] Official site of the «pandas» software product [Electronic Resource], URL: https://pandas.pydata.org (date of the request: 27.05.2020).

[8] Official site of the «numpy» software product [Electronic Resource], URL: https://numpy.org/ (date of the request: 31.05.2020).

[9] Logistic Regression Module sklearn [Electronic Resource], URL: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html (date of the request: 30.05.2020).

[10] Sklearn Decision Tree Classifier Module [Electronic Resource], URL: https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html (date of the request: 30.05.2020).

[11] Sklearn random forest classifier module [Electronic Resource], URL: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html#sklearn.ensemble.RandomForestClassifier (date of the request: 30.05.2020).

[12] Gradient Booster Classifier Module LightGBM [Electronic Resource], URL: https://lightgbm.readthedocs.io/en/latest/pythonapi/lightgbm.LGBMClassifier.html (date of the request: 30.05.2020).

[13] CatBoost Solution Tree Gradient Boost Classifier Module [Electronic Resource], URL: https://catboost.ai/docs/concepts/python-reference_catboostclassifier.html (date of the request: 30.05.2020)

[14] Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures [Electronic Resource], URL: https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/ (date of the request: 30.05.2020).

[15] Shap library module, [Electronic Resource], URL: https://shap.readthedocs.io/en/latest/, (date of the request: 03.06.20).

[16] bronchopulmonary_diag_system github repository [Electronic Resource], URL: https://github.com/EvgrafovAlexander/bronchopulmonary_diag_system (date of the request: 03.06.20).