


## DEVELOPING AN ARTIFICIAL INTELLIGENCE-BASED SYSTEM FOR MEDICAL PREDICTION

Sakhibgareeva MV , Zaozersky AYu

COMTEK LLC, Ufa, Russia

Diagnostic accuracy remains one of the central problems of medical care. In this work we attempt to apply artificial intelligence to solve this challenge. We propose an approach to medical prediction based on the intelligent analysis of patients' data from 200 different laboratory tests. The initial sample included 7, 918 cases falling into 4 nosological categories: D50 (iron deficiency anemia), E11 (non-insulin-dependent diabetes mellitus), E74 (other disorders of carbohydrate metabolism), and E78 (disorders of lipoprotein metabolism and other lipidemias), and was further divided into the training and testing datasets. Using gradient boosting, we constructed a machine learning model. The model demonstrated a recognition rate of 89 % (AUC-ROC) and a mean certainty in the diagnosis of 92 %. Our study proves feasibility of using machine learning in the analysis of this type of medical data. We are currently implementing a web-service for medical prediction as part of our *Healthcare* platform aiming at automation of clinical practice.

**Keywords:** artificial intelligence, analysis of medical data, machine learning, gradient boosting, laboratory diagnostics, nosological diagnosis, multiclass classification, iron deficiency anemia, lipidemia, carbohydrate metabolism disorders

 **Correspondence should be addressed:** Margarita Sakhibgareeva  
ul. Bekhtereva 16, kv. 48, Ufa, Russia, 450047; margarita.vl2011@gmail.com

**Received:** 23.11.2017 **Accepted:** 13.12.2017

## РАЗРАБОТКА СИСТЕМЫ ПРОГНОЗИРОВАНИЯ ДИАГНОЗОВ ЗАБОЛЕВАНИЙ НА ОСНОВЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

М. В. Сахибгареева , А. Ю. Заозерский

ООО «КОМТЕК», Уфа

В статье представлены результаты исследования по применению технологий искусственного интеллекта для решения одной из основных проблем здравоохранения — повышения качества диагностики заболеваний. Предложен подход к прогнозированию нозологических диагнозов путем интеллектуального анализа совокупности результатов лабораторных исследований (200 тестов), проводимых по каждому случаю заболевания пациентов. В общую выборку, разделенную впоследствии на обучающую и тестовую, включили данные о 7 918 случаях заболеваний по 4 нозологиям: D50 (железодефицитная анемия), E11 (инсулиннезависимый сахарный диабет), E74 (другие нарушения обмена углеводов), E78 (нарушения обмена липопротеидов и другие липидемии). Методом градиентного бустинга для них была построена модель машинного обучения. Точность распознавания моделью выбранных диагнозов составила более 89 % (ROC AUC) при средней уверенности модели в каждом прогнозируемом диагнозе в 92 %. Исследование показало принципиальную возможность применения методов машинного обучения для анализа данных такого рода. Система прогнозирования диагнозов заболеваний внедряется в виде веб-сервиса в программный комплекс «Здравоохранение», предназначенный для автоматизации работы медицинских учреждений.

**Ключевые слова:** искусственный интеллект, анализ медицинских данных, машинное обучение, градиентный бустинг, лабораторная диагностика, нозологический диагноз, многоклассовая классификация, железодефицитная анемия, липидемия, нарушения обмена углеводов

 **Для корреспонденции:** Сахибгареева Маргарита Владимировна  
ул. Бехтерева, д. 16, кв. 48, г. Уфа, 450047; margarita.vl2011@gmail.com

**Статья получена:** 23.11.2017 **Статья принята к печати:** 13.12.2017

Development of information technologies aimed to facilitate the efficient delivery of medical care is one of the priority goals set for the Russian healthcare system. Increasing effort is being made to improve the quality of healthcare through the use of information systems, expedite transition from paper files to electronic medical records and employ data mining for the analysis of huge arrays of medical data [1, 2].

Collection of medical data still presents a problem, as noted in a number of works [3, 4], which seriously impedes their digitalization necessary for machine learning and delays development of analytical software. Our close collaboration with the Siberian Center for Information Protection and deployment of the original *Zdravookhranenie* software in a few regional

medical centers allowed us to build a vast database of medical records and obtain authorization to process these data. It was a perfect opportunity to perform data mining using machine learning techniques.

The use of diagnostic information systems in clinical practice can be very beneficial for patients. High workload or the lack of expertise affects clinical decisions doctors make. Besides, a taking into account of a set of information about the patient is a basis for accurate diagnosis, prediction of disease progression and treatment planning; without it clinical decisions are mere approximations [5].

According to A. Chuchalin's report presented at the Second National Congress of GPs, every third case in Russia is

misdiagnosed [6]. Likewise, we have discovered a considerable number of diagnostic errors while analyzing the records of a few healthcare facilities that use our software. In the course of our analysis, we calculated the discrepancies between the definitive and preliminary diagnoses. Results are presented in Table 1 which features distribution of erroneous diagnoses across different departments of healthcare facilities and Table 2 showing the percentage of erroneous diagnoses in different nosological categories. Names of the healthcare facilities are not provided in this article for ethical reasons.

Not only patients becomes victims of wrong preliminary diagnoses and get useless treatments but also medical clinics incur considerable expenses: the Fund of Compulsory Health Insurance only subsidizes treatments based on a definitive diagnosis.

In view of this, we decided that prediction of nosological diagnosis should be a priority task in the development of an artificial intelligence-based system. The aim of this work was to test the feasibility of medical data mining using machine learning, to assess prediction accuracy that makes a machine learning model useful, and to enhance our *Zdravookhranenie* platform.

## METHODS

### Initial dataset

Medical decisions can be based on a medical history, physical examinations, and results of laboratory or complex functional tests. Lab tests provide the most objective information about patient's condition and are often used when other methods have failed to identify or confirm a pathology. These tests are especially useful in patients with anemia, lipidemia, hepatitis, seropositive rheumatoid arthritis, etc.

The source dataset consisted of disease cases with established definitive diagnoses. The feature space included patients' sex and age and the results of laboratory tests obtained from the data of prophylactic medical examination. The data were collected using our *Zdravookhranenie* software solution [7]. We chose 4 nosologies for the analysis, including D50, E11, E74, and E78, that can be suspected and diagnosed based on laboratory tests. The initial dataset was as follows:

- iron deficiency anemia (D50) — 778 cases (10 %);
- non-insulin-dependent diabetes mellitus (E11) — 1,392 cases (17 %);
- other disorders of carbohydrate metabolism (E74) — 163 cases (2 %);

- disorders of lipoprotein metabolism and other lipidemias (E78) — 5,585 cases (71 %).

In total, the dataset included 7,918 cases with results of 200 laboratory tests (blood and urine tests, cytologic examinations, etc.) that occurred during the period from 2005 to 2017 with patients aged 18 to 99 years, of whom 71 % of were females and 29 % were males. In some cases, the results of laboratory tests were recorded as “normal”, “below the norm” and “above the norm”.

### Choosing a method of machine learning and performance metrics

Prediction of diagnosis based on the results of laboratory tests is a multiclass classification problem.

The data were analyzed using Scikit-learn [8], a Python-based open-source library for machine learning. We carried out a few preliminary tests involving such methods of machine learning as neuronal networks, decision trees, and gradient boosting. The last one showed the best results for our problem. It is a technique in which an ensemble of predictors is built sequentially, with every subsequent algorithm compensating for the mistakes of a previous predictor [9]. Gradient boosting over decision trees is believed to be the most effective universal method of machine learning. Decision trees also perform very well in classification tasks.

Considering the specifics of the problem and the fact that the initial dataset was imbalanced, we selected performance metrics with special care. The metrics will be described below in terms of a confusion matrix [9-10] with respect to multiclass classification using the one-against-all approach. This approach is based on reducing multidimensional classifications to a set of binary tasks in which a picked class is classified as 1, and the rest classes are classified as 0. For every picked class  $i$  the following parameters are determined:

- TP (True Positive) — the number of true positive instances correctly assigned to class  $i$ ;
- TN (True Negative) — the number of true negatives instances correctly not assigned to class  $i$  and therefore assigned to class  $j \neq i$ ;
- FP (False Positive) — the number of false positives instances incorrectly assigned to class  $i$ ;
- FN (False Negative) — the number of false negatives instances incorrectly assigned to class  $j \neq i$  that should have been assigned to class  $i$ .

Accuracy is the most intuitive performance metric showing a fraction of correct responses; however, is not suitable for imbalanced datasets.

**Table 1.** Percentage of wrong diagnoses in different units of several healthcare agencies based in Russia

Unit	Percentage of wrong diagnoses. %		
	Healthcare agency 1	Healthcare agency 2	Healthcare agency 3
Pulmonary	76.80	39.28	–
Anaesthetics and Intensive care	72.96	24.11	73.11
Cardiac care (>1)	57.88	23.00	46.43
Therapeutic	56.36	–	–
Gastroenterology	66.38	11.29	–
Trauma	32.19	–	60.64
Neurology	55.04	14.97	–
Urology	–	–	67.72

**Table 2.** Percentage of wrong diagnosis per nosological category in several Russia-based healthcare agencies

Nosology	Percentage of wrong diagnoses, %
Disorders of lipoprotein metabolism and other lipidemias	92.73
Cholera	88.89
Disorders of sphingolipid metabolism and other lipid storage disorders	88.72
Immunodeficiency with predominantly antibody defects	83.33
Sequelae of other and unspecified infectious and parasitic diseases	80.00
Evidence of alcohol involvement determined by blood alcohol level	80.00
Juvenile arthritis in diseases classified elsewhere	75.00
Other bacterial diseases, not elsewhere classified	66.67
Car occupant injured in collision with pedal cycle	66.67
Lactose intolerance	60.00
Pericarditis in diseases classified elsewhere	60.00
Trichomoniasis	50.00
Other intestinal helminthiases, not elsewhere classified	50.00
Viral agents as the cause of diseases classified elsewhere	50.00
Malignant neoplasms of lip	50.00
Carcinoma in situ of cervix uteri	50.00
Deficiency of other nutrient elements	50.00
Other diseases of inner ear	50.00
Intestinal malabsorption	50.00
Hypertrichosis	50.00
Other disorders of kidney and ureter in diseases classified elsewhere	50.00
Pre-existing hypertension with pre-eclampsia	50.00
Epidermolysis bullosa	50.00
Unspecified jaundice	50.00
Anomaly of leukocytes, not elsewhere classified	50.00
Glycosuria	50.00
Other and unspecified abnormal findings in urine	50.00
Other disorders of carbohydrate metabolism	20.70
Iron deficiency anemia	13.90
Non-insulin-dependent diabetes mellitus	3.240

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

Therefore, other metrics are often used instead, including:

- precision — a fraction of true positives instances among all predicted positives. In other words, it shows how many positive predictions were really positive:

$$precision = \frac{TP}{TP + FP};$$

- recall — a fraction of true negatives instances among all true and false positives. It is also known as a true positive rate (TPR):

$$recall = \frac{TP}{TP + FN}.$$

Recall is used to evaluate performance of a machine learning model when there is a need to reduce the number of false negatives (FN) and measure all positives [10]. This metric is preferred for medical diagnostic tasks when it is important not

to miss a diagnosis. Although it is quite intuitive, it is not always good for imbalanced datasets.

Another metric used in our study was ROC AUC recommended in [10] for the evaluation of model performance on imbalanced datasets. AUC stands for area under [ROC] curve, ROC is receiver operating characteristic. This curve is constructed by plotting the true positive rate (TPR) against the false positive rate (FPR) and is a line connecting (0, 0) to (1, 1):

$$TPR = \frac{TP}{TP + FN},$$

$$FPR = \frac{FP}{FP + TN}.$$

It is believed that the higher the ROC AUC value, the better the performance of the classifier. ROC AUC of 0.5 means the classifier makes random guesses. ROC AUC below 0.5 means that the classifier does the opposite of what is expected of it: if true positives were labeled as negatives, it would perform better.

Considering the above said, we used ROC AUC as a primary metric, but also accounted for recall.

**Table 3.** Performance of the machine learning model designed for diagnostic prediction

Diagnosis	Metric				Dimension of the test set, number of cases
	ROC AUC	Recall	Precision	Accuracy	
D50 (iron deficiency anemia)	0.98	0.66	0.83	0.95	44
E11 (non-insulin-dependent diabetes mellitus)	0.91	0.62	0.69	0.9	69
E74 (other disorders of carbohydrate metabolism)	0.89	0.21	0.6	0.97	14
E78 (disorders of lipoprotein metabolism and other lipidemias)	0.94	0.96	0.89	0.89	318

## RESULTS

The diagnoses and the results of laboratory tests were divided into two sets: the training set (75 % of cases) and the test set (25 % of cases). The model was built for 4 nosological categories (D50, E11, E74, E78) using gradient boosting. For the test set ROC AUC was above 89 % (Table 3). Mean certainty in correct diagnoses included in a test sample was 92 %.

## DISCUSSION

High ROC AUC values falling between 89 % and 98 % indicate that our model is feasible for the prediction of the studied diagnoses. Importantly, our dataset consisted of various data types, including the results of 200 different laboratory tests and such parameters as patients' sex and age. Among other strengths of the study is the use of enough large dataset accumulated over the course of a few years. For example, in [11] the analysis was carried out on the data collected over the period of just 3 months in a Boston hospital. The authors of the study attempted to predict ferritin blood levels. They also used ROC AUC as quality metric which turned to be as high as 97%. However, it should be noted that according to a number of research works [12–14] a focus on nosological categories may increase prediction accuracy. According to [15, 16], performance can be enhanced through the use of different methods for medical data preprocessing.

## CONCLUSIONS

Our study has proved the feasibility of machine learning techniques for the analysis of our medical records. Currently, we are incorporating this model into our *Zdravookhranenie* software. We are working on a web service which will accumulate and analyze the results of all laboratory tests specified in a patient's medical history. The web service will "report" to the *Zdravookhranenie* platform the results of the analysis and returning the most probable diagnoses that a doctor may take into for appointment of a treatment regimen.

We are planning to include more nosologies into our model and improve its quality by designing separate models for each diagnosis. These models will account for the laboratory tests that affect the prediction outcome the most. Thus, we will be able to start developing a tool that can recommend the most relevant lab tests for the diagnosis of a particular condition.

We hope that our work will expedite transition to personal medicine [17, 18] based on the analysis of patient's unique medical records not limited to the results of the laboratory tests. This task can be solved using artificial intelligence for diagnostic prediction and generating personalized treatment recommendations. This will help to reduce the number of medical errors and increase clinical significance of prevention measures by monitoring patient's records.

## References

- Gusev AV. [Perspectives of neural networks, and deep machine learning to create solutions for healthcare]. Doctor and information technologies. 2017; (3): 92–105. Russian.
- Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. Artificial Intelligence in Medicine. 2001; 23(1): 89–109.
- Bledzhants GA, Sarkisian MA, Isakova IA, Tumanov NF, Popov AN, Begmurodova NS. [The key technologies of artificial intelligence in medicine]. Remedium. Magazine about the Russian market of medicines and medical equipment. 2015; (12): 10–5. Russian.
- [Machine learning helps physicians to make more informed decisions]. Telemedicina.ru [Internet]. 2017 Sep. [cited 2017 Sep 4]. Available from: <https://telemedicina.ru/news/equip/mashinnoe-obuchenie-pomojet-vracham-prinimat-bolee-informirovannyye-resheniya>. Russian.
- Zharikov OG, Meshcherikov IV, Litvin AA. [Neuronet technologies in medicine]. The issues of organization and Informatization of healthcare. 2007; 4 (53): 59–63. Russian.
- Golovachev V. Oshibochniy diagnoz. Trud. 2014 Oct 28. Russian.
- Korotaev IG, Chernukhin GA, the authors; COMTEK Ltd., assignee. Software complex «Healthcare». The certificate of official registration program for computer 2007613347. 2007 Aug 9.
- Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011; 12: 2825–30.
- Shalev-Shwartz S, Ben-David S. Understanding Machine Learning: From Theory to Algorithms. New York: Cambridge University Press; 2014. 410 p.
- Müller AC, Guido S. Introduction to Machine Learning with Python: A Guide for Data Scientists. 1st ed. O'Reilly Media; 2016. 285 p.
- Luo Y, Szolovits P, Dighe AS, Baron JM. Using Machine Learning to Predict Laboratory Test Results. Am J Clin Pathol. 2016 Jun; 145 (6): 778–88. DOI: 10.1093/ajcp/aqw064.
- Khlivnenko LV, Piatakovich FA. [The option of constructing the system of collective human-machine intelligence for big data processing in medicine]. Health and Education Millenium. 2016; 18 (12): 141–4. Russian.
- Bilenko AA, Rybkin SV. [The application of machine learning algorithms to identify high risk diabetes type 1 diabetes]. E-magazine: science, technology and education. 2017; 1 (10): 44–9. Russian.
- Tseng CJ, Lu CJ, Chang CC, Chen GD, Cheewakriangkrai C. Integration of data mining classification techniques and ensemble learning to identify risk factors and diagnose ovarian cancer recurrence. Artif Intell Med. 2017; (78): 47–54.
- Oniško A, Druzdzel MJ. Impact of precision of Bayesian network parameters on accuracy of medical diagnostic systems.

- Artif Intell Med. 2013 Mar; 57 (3): 197–206. DOI: 10.1016/j.artmed.2013.01.004.
16. Khajehali N, Alizadeh S. Extract critical factors affecting the length of hospital stay of pneumonia patient by data mining (case study: an Iranian hospital). *Artif Intell Med.* 2017 Nov; 83: 2–13. DOI: 10.1016/j.artmed.2017.06.010.
  17. Weiss JC, Natarajan S, Peissig PL, McCarty CA, Page D. Machine Learning for Personalized Medicine: Predicting Primary Myocardial Infarction from Electronic Health Records. *AI Magazine.* 2012; 33 (4): 33–45.
  18. Futoma J, Sendak M, Cameron B, Heller K. Predicting Disease Progression with a Model for Multivariate Longitudinal Clinical Data. In: *Proceedings of the 1st Machine Learning for Healthcare Conference*; 2016 Aug 19–20; Children's Hospital LA, USA; 2016; (56): 42–54.

## Литература

1. Гусев А. В. Перспективы нейронных сетей и глубокого машинного обучения в создании решений для здравоохранения. *Врач и информационные технологии.* 2017; (3): 92–105.
2. Kononenko I. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine.* 2001; 23(1): 89–109.
3. Бледжянц Г. А., Саркисян М. А., Исакова Ю. А., Туманов Н. Ф., Попов А. Н., Бегмуродова Н. Ш. Ключевые технологии формирования искусственного интеллекта в медицине. *Ремедиум. Журнал о российском рынке лекарств и медицинской технике.* 2015; (12): 10–5.
4. Машинное обучение поможет врачам принимать более информированные решения. *Телемедицина.ru* [Интернет]. Сентябрь 2017 г. [цитировано 4 сентября 2017 г.]. Доступно по ссылке: <https://telemedicina.ru/news/equip/mashinnnoe-obucheniye-pomojet-vracham-prinimat-bolee-informirovannyye-resheniya>.
5. Жариков О. Г., Мещеряков Ю. В., Литвин А. А. Нейросетевые технологии в медицине. *Вопросы организации и информатизации здравоохранения.* 2007; 4 (53): 59–63.
6. Головачев В. Ошибочный диагноз. *Газета «Труд».* 28 октября 2014 г.
7. Коротаяев И. Г., Чернухин Г. А., авторы; ООО «КОМТЕК», правообладатель. Программный комплекс «Здравоохранение». Свидетельство об официальной регистрации программы для ЭВМ № 2007613347 от 09.08.2007.
8. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research.* 2011; 12: 2825–30.
9. Shalev-Shwartz S, Ben-David S. *Understanding Machine Learning: From Theory to Algorithms.* New York: Cambridge University Press; 2014. 410 p.
10. Müller AC, Guido S. *Introduction to Machine Learning with Python: A Guide for Data Scientists.* 1st ed. O'Reilly Media; 2016. 285 p.
11. Luo Y, Szolovits P, Dighe AS, Baron JM. Using Machine Learning to Predict Laboratory Test Results. *Am J Clin Pathol.* 2016 Jun; 145 (6): 778–88. DOI: 10.1093/ajcp/aqw064.
12. Хливненко Л. В., Пятакович Ф. А. Вариант построения системы коллективного человеко-машинного интеллекта для обработки больших данных в медицине. *Здоровье и образование в XXI веке.* 2016; 18 (12): 141–4.
13. Биленко А. А., Рыбкин С. В. Применение алгоритмов машинного обучения для определения высокого риска сахарного диабета 1 типа. *Электронный журнал: наука, техника и образование.* 2017; 1 (10): 44–9.
14. Tseng CJ, Lu CJ, Chang CC, Chen GD, Cheewakriangkrai C. Integration of data mining classification techniques and ensemble learning to identify risk factors and diagnose ovarian cancer recurrence. *Artif Intell Med.* 2017; (78): 47–54.
15. Oniśko A, Druzdzel MJ. Impact of precision of Bayesian network parameters on accuracy of medical diagnostic systems. *Artif Intell Med.* 2013 Mar; 57 (3): 197–206. DOI: 10.1016/j.artmed.2013.01.004.
16. Khajehali N, Alizadeh S. Extract critical factors affecting the length of hospital stay of pneumonia patient by data mining (case study: an Iranian hospital). *Artif Intell Med.* 2017 Nov; 83: 2–13. DOI: 10.1016/j.artmed.2017.06.010.
17. Weiss JC, Natarajan S, Peissig PL, McCarty CA, Page D. Machine Learning for Personalized Medicine: Predicting Primary Myocardial Infarction from Electronic Health Records. *AI Magazine.* 2012; 33 (4): 33–45.
18. Futoma J, Sendak M, Cameron B, Heller K. Predicting Disease Progression with a Model for Multivariate Longitudinal Clinical Data. In: *Proceedings of the 1st Machine Learning for Healthcare Conference*; 2016 Aug 19–20; Children's Hospital LA, USA; 2016; (56): 42–54.