Using machine learning methods in problems with large amounts of data

Olga Kuimova[1,*], *Vladislav* Kukartsev[1,2], *Artem* Stupin[1], *Ekaterina* Markevich[2] and *Stanislav* Apanasenko[1]

[1]Reshetnev Siberian State University of Science and Technology, 31, Krasnoyarsky Rabochy Av., 660037 Krasnoyarsk, Russian Federation
[2]Siberian Federal University, 79, Svobodny Av., Krasnoyarsk, 660041, Russian Federation

Abstract. This article explores the use of artificial intelligence in medicine, in particular in radiology, pathology, drug development. The usefulness of robotic assistants in the medical field is revealed, including machine learning in medical science, as well as routing in hospitals. It also discusses such machine learning methods as classification methods, regression restoration methods, clustering methods. As a result, based on what is considered in this article, it is concluded that manual processing becomes more complicated and impossible with a large amount of data. There is a need for automatic processing that can transform modern medicine. And also, conclusions were made about how accurately the deep learning mechanisms can provide a more accurate result in the processing and classification of images compared to the results obtained at the human level. It became clear that deep learning not only aids in the selection and extraction of characteristics, but also has the potential to measure predictive target audiences and provide proactive predictions to help clinicians go a long way.

1 Introduction

In the modern world, great attention is paid to the development of computer technology. Every day, thousands or even millions of lines of code come out from under the fingers of programmers. Smart homes, smart cars, convenient electronic payment systems, online banking, entertainment services - all this would be impossible without writing code. However, the software capabilities are not limited to these examples only. While some technologies are designed to make human life easier in everyday life, others serve to preserve, and sometimes even save human life. The purpose of this article is to provide a brief excursion into the world of medicine and, in particular, into the world of technologies used in this area [1-3].

It's no secret that healthcare is one of the most important development niches at all times. And, like any other area, it undoubtedly needs modernization, and more importantly, optimization. Throughout the history of medicine, an unimaginable number of diseases have been studied, as well as ways to treat them. However, as practice shows, with the development of information technologies and over time, the database of this data is constantly being replenished. Moreover, more common issues often arise, for example, regarding the placement of patients or scheduling of appointments. Let's not forget the fact that the correct and timely diagnosis is constantly playing a leading role in the patient's recovery. After all, the process of training specialists is also extremely labor-intensive and resource-intensive [4-6].

From all of the above, an intermediate conclusion can be made that an exorbitant burden falls to the shoulders of a person, even the most successful in the field of medicine, which is almost impossible to cope with alone. This is where modern technologies come into play, designed to facilitate the work of professionals and help save lives. One of the most optimal solutions is the use of artificial intelligence technologies, in particular, the use of trained neural networks. The use of this technology makes it possible to facilitate the work of specialists in making even the most complex diagnoses, to increase the accuracy of detecting diseases, and also to solve a number of optimization issues that may arise in medical organizations [1]. So, neural networks are able to provide assistance not only in a quick and correct diagnosis, but also in solving more everyday issues related to the internal "kitchen" of medical institutions [7-9].

Further in the article we will give examples of the use of ML technologies directly in medicine, ML methods used to train neural networks, as well as the tasks solved by these neural networks [10-12].

2 Materials and Methods

2.1. The use of artificial intelligence in medicine

* Corresponding author: olga-barishnikova@yandex.ru

As mentioned earlier, the use of neural networks allows you to solve a number of medical problems that require immediate solution. One way or another, the most obvious and common issue solved by the use of neural network technologies is the diagnosis. So, today there are many examples of the use of this technology for the benefit of more accurate and effective medicine [5].

- AI in radiology.

Machines and algorithms have shown themselves to be effective in the classification of images, as well as in the recognition of objects in the photo. Hence, it naturally follows that they can effectively interpret imaging data, providing a wide range of possible applications of artificial intelligence algorithms, in particular in the field of radiology.

- AI in pathology.

Another interesting direction in the development of medicine is the development of an algorithm to assist a specialist.

Harvard Medical School is developing DeepLearning-based technology capable of detecting tissue pathologies. Combined with human experience, this technology can greatly enhance research efficiency by detecting pathological changes that the human eye may not be able to track. This technology in combination with humans has already shown excellent accuracy results (99.5%).

- Drug development.

The trend of using machine learning in medical research is quite popular. For example, the use of AI technologies can solve the following problem.

The experience of pharmaceutical companies shows that it takes approximately 12 years from the start of preclinical trials to the approval of a drug and treatment of patients. At the same time, only 0.1% of “candidate drugs” go to clinical tests. Approval is received by 20% of them.

A solution from San Francisco-based AtomWise, AtomNet uses all of the same deep learning technologies. A giant dataset of data on the interactions of molecules already known to science served as an excellent basis for training a neural network. The result was an algorithm that predicts the behavior of certain particles that interact. This technology can already be considered successful, as it helped to develop a cure for Ebola.

The first technology is surgical robot assistants.

Already today, many operations are carried out using machine vision technologies and human-controlled manipulators [6]. This combination is great for conventional surgeons. For example, this technology is able to "insure" the surgeon against accidents and inattention. Moreover, there is a clear advantage in improving the surgeon's visibility and reminding him of the order of the operation. In addition, the AI will assist in the selection of the tissue incision and suture to be applied, which will help alleviate the patient's pain during the recovery process.

The second is machine learning and medical science.

Departing a little from the topic of direct practical application of neural networks in diagnostics and operations, it is worth mentioning such an important problem of modern medicine as data structuring.

All medical facilities generate petabytes of various data about patients, their diseases and recovery processes. This data is unfortunately chaotic and unstructured. An additional complication is that, unlike, for example, business data, medical databases do not lend themselves well to traditional analytical methods.

A powerful AI-enabled platform is able to efficiently analyze information, find patterns and, therefore, create a certain structure that describes the available data. Thanks to this, it becomes possible to form more accurate data on the patient's health, which is an extremely important aspect in the field of medicine.

The third is routing in hospitals.

Another common, but no less important problem that can be solved by ML methods is routing in hospitals. This means the following. When a patient is admitted to the hospital, the medical staff carries out standard procedures: taking blood for analysis, taking anamnesis, etc. In 80% of cases, the patient is prescribed medication and sent home.

University College London Bloomsbury Clinic uses artificial intelligence systems to identify the very 20% of patients in urgent need of care. Based on standard analyzes, the system will give priority to one or another patient whose health is or will be under serious threat.

Scientists at the National Hospital of Neurology and Neurosurgery in the UK have developed a machine learning algorithm that analyzes information about appointments at a clinic and estimates the likelihood that a patient, for one reason or another, will miss an MRI scan. Their system considers such parameters as the age of a person, his address and distance to the clinic, weather conditions.

And scientists from the University are developing a system that monitors the "movement" of medical personnel and patients in the hospital: the approximate location according to the schedule, attendance of procedures, etc. Based on this data, the system will be able to determine situations or places where there is a potential shortage of equipment or the accumulation of queues.

This is not a complete list of the possibilities of applying machine learning technologies in medicine, but it shows the main branches of the development of this area, as well as the importance of developing and implementing software based on machine learning in medical organizations.

2.2 Machine learning methods

Machine learning is based on the idea that analytic systems can learn to identify patterns and make decisions with minimal human input. Machine learning is a class of artificial intelligence methods that are aimed not at solving a problem directly, but at learning in the process of solving similar problems. Machine learning focuses on developing computer programs that have ways to process data and learn on their own. This process is accompanied by the study of data (examples,

instructions) in order to identify relationships or patterns in order to apply the acquired “knowledge” to make decisions or make predictions. The system "trains" in order to improve the accuracy of its forecasts and decisions [4].

The result directly depends on the choice of a machine learning method, since each method is most effective for a certain range of tasks. In addition, different methods use different resources, so some methods may be more cost-effective in a particular situation. In medicine, where no mistakes can be made, and performance should strive for the maximum level, the choice of teaching method is especially important.

Approaches to automatic learning can be divided into at least two types: with a teacher and without a teacher [7]. In the first case, the system receives training cases, each of which is represented by a set of input values along with the correct answer for it. The goal is to acquire the ability to respond appropriately to new cases. Supervised learning tasks include classification and regression reconstruction tasks.

2.3 Classification problem

The classification task is a formalized task where there are many objects (situations), divided into classes according to some rule. In the problem, a finite set of objects is given, for which it is known to which classes they belong. The set is commonly referred to as a sample. The class of the remaining objects is unknown. The classification problem is that it is required to build an algorithm capable of classifying an arbitrary object from the original set [8].

The phrase "classify an object" is usually understood to indicate the number or name of the class to which this object belongs. Classification is one of the sections of machine learning. The classification problem in machine learning falls under the heading of supervised learning.

There are various types of input data for the classification problem. Let's list some of them and reveal their essence:

• Feature description is one of the most common cases of classification problems. Each object in a class is described by a set of its specific characteristics, called attributes. Signs are both numeric and non-numeric.

• Matrix of distances between objects - each object in the class is described by the distance to all other objects of the training sample.

• Time series or signal is a sequence of measurements in time. Any dimension can be represented by a number or a vector, and sometimes by an indicative description of the object under study at a certain point in time.

• Video sequence or image, for example, X-rays of patients.

• Graphs, texts, results of database queries, etc. are perhaps the most difficult cases of processing input data. Usually they try to bring them to the first or second case by means of preliminary data processing and feature extraction.

There are also different types of classes that are included in the main typology of classification problems:

• Two-class classification,
• Non-overlapping classes,
• Multi-class classification,
• Fuzzy classes,
• Overlapping classes.

2.4 Classification methods

In the medical field, when constructing diagnostic systems, classification methods are often used, such as the naive Bayesian classifier (NaiveBayesClassifier), logistic regression (Logistic regression), decision trees (DecisionTrees), support vector machine (SupportVectorMachine) [9].

NaiveBayes consists of two types of probabilities that are calculated using the training data:

- probability of each class;
- conditional probability for each class for each value of x.

After calculating the probabilistic model, it can be used to make predictions with new data using Bayes' theorem. If you have real data, then, assuming a normal distribution, it is not too difficult to calculate these probabilities. Naive Bayes is called naive because the algorithm assumes that each input variable is independent. This is a strong assumption and does not match the actual data. Nevertheless, this algorithm is very effective for a number of complex tasks. For example, Naive Bayes is often used to classify text, filter spam, and generate recommendation systems [10].

SVM (Support Vector Machine) is a set of similar supervised learning algorithms used for classification and regression analysis problems. Belongs to the family of linear classifiers. A special feature of the support vector machine is that the empirical classification error continuously decreases and the gap increases, which is why the method is also known as the maximum gap classifier method.

The main idea of the method is to translate the initial vectors into a space of higher dimension and search for a separating hyperplane with a maximum gap in this space. Two parallel hyperplanes are constructed on both sides of the hyperplane separating the classes. The separating hyperplane is the hyperplane that maximizes the distance to two parallel hyperplanes [12]. The algorithm works under the assumption that the greater the difference or distance between these parallel hyperplanes, the smaller the average error of the classifier will be. This method is often used for face recognition, text categorization, handwriting recognition.

2.5 Regression reconstruction problem

The task of restoring regression is to find the approximate dependence of one group of values on another group of values. When solving the problem of restoring regression, the input data and the corresponding output values are analyzed in order to derive an approximate pattern, which will then be used to determine the output value with the new input data. This problem is encountered, for example, when

assessing the value of real estate: by the size of the apartment, the height of the ceilings, the welfare of the area, the distance from the metro, etc. it is necessary to estimate the cost of housing. Also, this problem occurs, for example, when assessing mortality in certain injuries or when determining the healing time of an organ based on postoperative indicators [11].

2.6 Regression recovery methods

In the medical field, when constructing diagnostic systems, such regression methods as linear regression methods (Linear Regression), nonlinear regression methods (Nonlinear Regression), decision trees (DecisionTrees), and support vector method (SupportVectorRegression) are often used.

LinearRegression (linear regression) is one of the methods of regression analysis, which is used when there is a linear relationship between the set of independent variables (factors, regressors) and the dependent variable. In this case, we have a linear combination of the given basis functions, and the goal of the regression is to find the coefficients of this linear combination, while minimizing the loss function. This method is used in forecasting sales volumes, forecasting the value of securities, and analyzing the elasticity of demand.

DecisionTreeRegression is a method that builds a model using a tree structure. This method splits the dataset into smaller and smaller subsets while developing an associated decision tree. This method uses an algorithm that reduces the standard deviation [8].

2.7 Clustering problem

In unsupervised learning, the machine is not provided with outputs, only inputs, so it must try to find meaningful information on its own. An example of an unsupervised learning problem is the clustering problem. This task consists in the need to group a set of objects into subsets (clusters) in such a way that objects from one cluster are more similar to each other than to objects from other clusters according to some criterion. This task is often encountered in economic geography (dividing countries by indicators into groups with a similar economic situation), marketing (highlighting characteristic groups of consumers according to the degree of interest in a product), sociology (analyzing sociological surveys), medicine (identifying patterns of antibiotic resistance, highlighting various types of fabrics in the 3D image).

2.8 Clustering methods

The most popular clustering methods are the k-means method and DBSCAN.

K-means (k-means method) is the most popular clustering method. The operation of the algorithm is such that it seeks to minimize the total square deviation of cluster points from the centers of these clusters. This algorithm splits the set of elements into a predetermined number of clusters k. The main idea is that at each iteration, the center of mass is recalculated for each cluster obtained in the previous step. This method is used in marketing when dividing into groups by behavior (purchase history, activity in the application), identifying anomalies, categorizing inventory [1].

DBSCAN (Density-Based Spatial Clustering for Noisy Applications) is a density-based data clustering algorithm - if given a set of points in some space, then the algorithm groups together points that are closely spaced), marking as outliers the points that are lonely in areas with low density.

3 Results and Discussion

Artificial neural networks are an attempt to model the information retrieval capabilities of nervous systems. The main difference between neural networks and conventional computing systems lies in the massive parallelism and redundancy that they use to cope with the unreliability of individual computing units. At the moment, there are a large number of neural network architectures, but there are several main ones that are used most often.

RNN (Recurrent Neural Network) is one of the fundamental network architectures that other deep learning architectures are built on top of. RNNs can use their internal state (memory) to process variable-length input sequences [7]. Each processed information is intercepted, stored and used to calculate the final result. This makes them useful for speech recognition, for example. Moreover, a recurrent network can have connections that refer to previous layers (or even the same layer). This feedback allows them to retain a memory of past entries and solve problems over time. RNNs in medicine can be used to make a diagnosis based on data (symptoms), or they also compete with CNNs in the task of recognizing pathologies in images.

Artificial neural networks are an attempt to model the information retrieval capabilities of nervous systems. The main difference between neural networks and conventional computing systems lies in the massive parallelism and redundancy that they use to cope with the unreliability of individual computing units. At the moment, there are a large number of neural network architectures, but there are several main ones that are most often used.

RNN (Recurrent Neural Network) is one of the fundamental network architectures that other deep learning architectures are built on top of. RNNs can use their internal state (memory) to process variable-length input sequences. Each processed information is intercepted, stored and used to calculate the final result. This makes them useful for speech recognition, for example. Moreover, a recurrent network can have connections that refer to previous layers (or even the same layer). This feedback allows them to retain a memory of past entries and solve problems over time. RNNs in medicine can be used to make a diagnosis based on data (symptoms), or they also compete with CNNs in the task of recognizing pathologies in images [11].

According to the research conducted, the method of convolutional neural networks shows high efficiency when applied in the medical field. For medical image classification problems, this method shows higher accuracy than methods such as SupportVectorMachine or ArtificialNeuralNetworks. However, in other tasks, such as device design, methods such as SupportVectorMachine, RandomForest, or perceptron are used more often and show high efficiency. This method is often used to examine images in radiology.

A DBN (Deep Web of Trust) is a multi-layered network (usually deep, including many hidden layers) in which each pair of connected layers represents a restricted Boltzmann machine (Restricted Boltzmann Machine (RBM)). A DBN consists of many layers of hidden variables ("hidden units"), with connections between layers, but not between units within each layer. Unlike other models, each layer in the DBN network learns the entire input signal. In CNN, the first layers filter inputs based on basic characteristics only, while the second layers recombine all the simple patterns found by the previous layers. DBNs operate in a holistic manner and regulate each layer in order.

The operation of such a network is rather complicated and is rarely used in medicine. But it is still used in some cases, for example, to study the progression of a poorly understood disease.

DSN (Deep Stacking Network) architecture is different from others. DSN is also often referred to as DCN-DeepConvexNetwork. The DSN / DCN includes a deep learning network, but it is actually a collection of separate deep learning networks. Each network in a DSN has its own hidden layers that process data [3]. This architecture was designed to improve on the learning problem, which is quite complex when it comes to traditional deep learning models. Due to its many levels, DSN considers learning not the only problem to be solved, but a set of individual problems.

Networks with such an architecture are used as a rule to diagnose various diseases based on indirect indications. Thus, it is possible to determine, for example, the presence of diabetes mellitus in a person without taking a blood test.

5 Conclusions

The field of medical imaging is gaining importance with the requirement for accurate and efficient diagnosis of disease within a short period of time. As manual processing becomes more complex, stagnant and impossible with a large amount of data, there is a need for automatic processing that can transform modern medicine. Deep learning mechanisms can provide higher accuracy in image processing and classification compared to results obtained at the human level. In-depth learning not only assists in the selection and extraction of characteristics, but also has the potential to measure predictive target audiences and provide proactive predictions to help clinicians greatly. Machine learning cannot replace doctors, but it can help streamline routine tasks, increase forecast accuracy, and simplify solving atypical problems [12].

References

1. N. Stylianou, A. Akbarov, E. Kontopantelis, I. Buchan, K.W. Dunn, *Burns* **41**, 925 (2015)
2. Y. Xu, Y. Wang, J. Yuan, Q. Cheng, X. Wang, P.L. Carson, *Ultrasonics* **91**, 1 (2019)
3. T. Maruyama, N. Hayashi, Y. Sato, S. Hyuga, Y. Wakayama, H. Watanabe, T. Ogura, *J. of X-ray Sci. and Tech.* **26**, 885 (2018)
4. Y.C. Han, K.I. Wong, I. Murray, *IET Signal Proc.* **14**, 243 (2020)
5. I. Banerjee, Y. Ling, M.C. Chen, S.A. Hasan, C.P. Langlotz, N. Moradzadeh, M.P. Lungren, *Art. intell. in med.* **97**, 79 (2019)
6. Y. Acikmese, S.E. Alptekin, *Pro. Com. Sci.* **159**, 658 (2019)
7. Y. Gautam *ISA transactions* (to be published)
8. M. Nilashi, H. Ahmadi, A. Sheikhtaheri, R. Naemi, R. Alotaibi, A.A. Alarood, A. Munshi, T.A. Rashid, J. Zhao, *ESA*. **159**, 113562 (2020)
9. Q. Zhang, J. Zhou, B. Zhang, E. Wu, *Inf. Sci.* **547**, 945 (2021)
10. A.A. Boyko, V.V. Kukartsev, K.Y. Lobkov, A.A. Stupina, *Jour. of Phys.: Conf. Ser.* **1015**, 042006 (2018)
11. V.V. Bukhtoyarov, V.S. Tynchenko, E.A. Petrovsky, V.V. Kukartsev, A.I. Kuklina, *Jour. of Phys.: Conf. Ser.* **1118**, 012041 (2018)
12. A. Milov, V. Tynchenko, V. Bukhtoyarov, V. Tynchenko, V. Kukartsev, *Proc. of the Com. M.in Sys.and Soft.* **1295**, 480 (2020)