Journal of Applied Engineering Science

doi:10.5937/jaes16-19634 Paper number: 16(2018)4, 563, 553 - 560

REVIEW OF THE BIG DATA TECHNOLOGY USE IN THE MEDICAL PROGNOSIS

Igor Koltunov[1*], Anton V. Panfilov[2], Ivan A. Poselsky[1], Nikolay N. Chubukov[2], Stanislav S. Matkov[2]

[1]Moscow Polytechnic University, Moscow, Russian Federation

[2]Tradition group LTD, Moscow, Russian Federation

The article shows the main aspects and problematics of elaborating effective models of current diagnostics and diagnostic prognosis of the patient's health status, who is an object of non–invasive monitoring, based on the current analysis of characteristic combinations of his/her vital signs on nosology and the results of long–term collecting, processing and semantic classificating the biomedical data.

Key words: Biosensor; Biosensor platform; Diagnostic informativeness; Diagnostic prognosis; Non–invasive monitoring; Nosology; Cloud technologies; Telemedicine; Health

INTRODUCTION

At present the big data technology is widely used in the sphere of medicine and health care. The section related to the big data includes the methods for storing, analyzing and processing a large amount of heterogeneous, structured and unstructured information. Since the information technology introduction in the health care sector, the medical industry has been storing and accumulating large amounts of data, such as observation records, medical histories, X–ray and tomographic images, insurance data and others. Using this data promises to bring many benefits in such areas as clinical decision making support, monitoring of the diseases, identifying critical trends and forecasting tendencies that affect public health.

The present article is dedicated to the analysis of successful applications of the big data technology in medicine and to the determination of prospects for the nearest development of such technologies.

The next section summarizes the features of the big data methodology. Working with a large data arrays using standard software and hardware tools is inefficient in terms of time expenditure and computational resources [1]. It is not only due to the amount of information, but also due to its diversity. The following is an overview of a number of significant results obtained, in particular, using specific aspects of the big data methodology. The conclusion presents the main findings and results.

THE BIG DATA POSSIBILITIES IN MEDICINE

The big data is a concept formed in information technologies at the beginning of this century. In addition to the obvious from the very term of a large amount of data, this concept implies its diversity of content and presentation forms, the absence of a formalized structure and systematization.

A variety of information provides the possibility to search for predicted patterns that are not visible when considering separated data segments. However, traditional data processing tools do not allow to effectively extract significant information from such sources. That is why the concept of the big data was formed, the systems and working methods were developed. Use of these technologies has led to many remarkable results in sociology and marketing.

Generally speaking, the big data methodology very timidly penetrates into medicine. The data associated with medicine may come from different sources: observation records, clinical decision support systems, government sources, laboratory data, etc. An important source of data may become "wearable devices" [02, 03, 04], which occupy a significant place in medical technology [05].

The physician can only focus on a small amount of information about the patient, which is collected in medical records or obtained as a result of a purposeful examination. Without special means of processing and analyzing the accumulated data the physician, when making a decision, is able to cover only a small part of even this information.

In this regard, the prospect of applying the big data methodology in medicine is very extensive, and the applications require research [06].

THE BIG DATA STORAGE ARCHITECTURE AND PROCESSING SYSTEM

The main difference between the standard data analytic methods and the big data analytic methods is that the big data technology performs a distributed calculation of this data. Huge arrays of information are broken into smaller ones and simultaneously processed on different machines, and finally the system produces an adequate processing result, which may be used to solve various tasks. Thus, the user is able to conduct an effective analysis of a large amount of information, including in a real–time environment [07]. Figure 1 illustrates the basic concept of the big data analysis system architecture.

Bolshaya Semenovskaya 38, 107023 Moscow, Russian Federation cposelskiyi@bk.ru

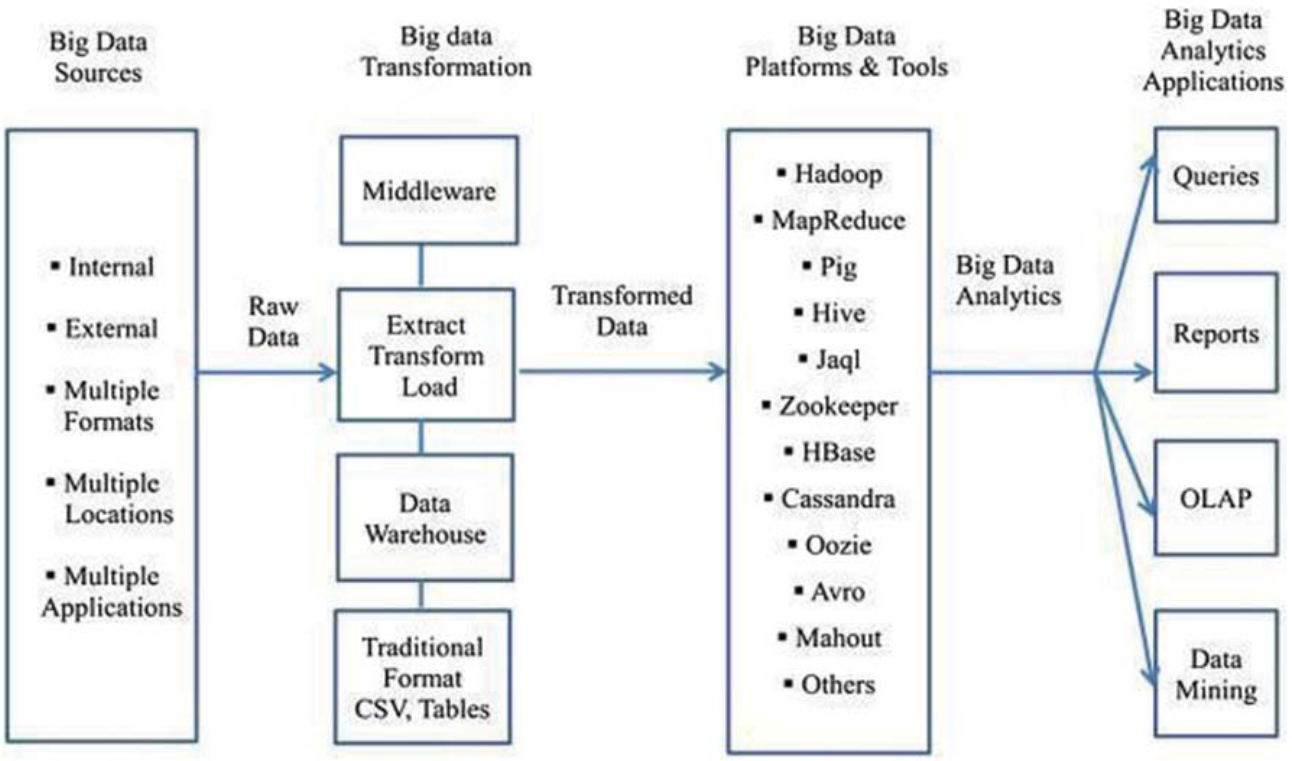

Figure 1: The concept of the big data analysis system architecture [06]

In the second step, the data is converted into one or more suitable formats for standard processing.

Further, the data is processed in accordance with the accepted analytical models and tools.

In the last step, the data is ready for display in the form of reports, queries, data mining and OLAP. These forms of presentation allow to further manipulate, aggregate and display the output data and to use it to solve applied tasks.

As a rule, the data is processed by such systems in delayed mode. But for tasks that require immediate recognition, for example, monitoring indicators in real time, Apache Spark technology is used, which replaces MapReduce in the above scheme.

THE BIG DATA PROSPECTS IN MEDICINE

The effectiveness of the big data technology introduction in the medical sector includes an increase in the quality of disease diagnostics, the accuracy of disease recognition in order to prevent them, and the study of these diseases.

In general, the big data has found application in the following health care sectors [06]:

- Clinical trials. Analysis of the clinical trials results of the medicinal product and the corresponding records of the patient taking this medicinal product are used to identify the side effects before the product appears on the market;
- Public health care. Tracking course of the disease and its frequency in a certain area is completed to preserve the level of public health;
- Vital signs monitoring. Acquisition and analysis of the real–time data from wearable sensors is performed in order to monitor the health status and to predict the risks and complications of a disease;
- Patient profile analysis. Study of the patient data is performed to find the optimal prevention and treatment.

REVIEW OF THE BIG DATA APPLICATIONS IN MEDICINE

Despite the fact that methodology of the big data is only at the initial stage of penetration into medicine, in recent years a number of practically significant results have been obtained based on methods close to it.

This section discusses some of the published results obtained using the big data methods. For each of them, the focus and key aspects of the study were determined. The results are presented in the form of 2 conventional groups: diagnostics of the disease and determination of the condition.

DISEASE RECOGNITION

Recognition of the Coronary Artery Problems

A research group from Boynor University, Iran, proposed a method for recognizing coronary insufficiency based on the neural network use in conjunction with a genetic algorithm to optimize balance. Also, to assess the quality of the proposed method, the recognition was performed without using a genetic algorithm. The support vector method was applied for extraction of signs. The classifi-

Journal of Applied Engineering Science Vol. 16, No. 4, 2018
ISSN 1451-4117

cation was carried out by means of a three–layer neural network [08].

Z–Alizadeh Sani dataset [09] was chosen as the data for training. It contains information about patients suffering from coronary insufficiency. It covers demographic characteristics, symptoms, examination results, electrocardiograms, echocardiograms, and laboratory tests. The results of the algorithm are presented in Table 1.

Table 1: Comparison of the test results. ACC=accuracy, SEN=sensitivity, SPEC=specificity

	SEN %	SPEC	ACC %	FPR	TPR
Neural network+genetic algorithm	97	92	93.85	0.08	0.97
Neural network	86	83	84.62	0.17	0.86

Recognition of the Myocardial Infarction Using a Convolutional Neural Network

In this study, an 11–layer convolutional neural network was used, that recognizes normal heartbeat and deviates from the norm [10]. This method is significant for the fact, that it can work under noisy data condition, since it works immediately with all signs, without selecting them and not carrying out any preprocessing.

The data is taken from the open source PTBDB [11]. This database contains data on 200 patients (148 patients and 52 healthy persons). For each patient, electrocardiogram data was recorded from a device with 12 electrodes. For testing dataset was divided into 2 parts, containing and not containing noise.

The number of iterations for testing is 60. The sampling was divided into training (90 %) and testing (10 %) parts. Cross–validation in 10 blocks was applied for splitting. Neural network characteristics:

- Batch size: 10;
- Coefficient of regularization: 0.2;
- Moment: 3 x 10 ^ (–4)
- Learning rate: 0.7.

The recognition of the results is presented in Table 2.

Table 2: Comparison of the test results. ACC=accuracy, SEN=sensitivity, SPEC=specificity

	TP	TN	FP	FN	ACC %	PPV %	SEN %	SPEC %
With noise	37.655	9.790	756	2.527	93.53	98.03	93.03	92.83
Without noise	38.368	9.933	613	1.814	95.22	98.43	95.49	94.19

Recognition and Localization of the Myocardial Infarction

For recognition and localization of the MI, the nearest neighbor method was applied [12]. This method involves recognizing the MI without learning. The idea is that the distance to the nearest measurement is calculated for each measurement. This algorithm is resource–intensive and can process data during a very long time. To reduce the recognition time the input data pruning (Arif–Fayyaz pruning) was applied. PTBDB [11] dataset was taken as data for training. It contains information about patients, recorded in real time (1000 measurements per second), from which noises were previously removed. The results of various types of the myocardial infarction recognition are presented in Table 3.

Table 3: Test results

Type	Sensitivity	Specificity
Anterior	99.5±0.13	99.09±0.25
Anterio–lateral	99.3±0.26	98.56±0.25
Anterio–septal	99.2±0.08	99.48±0.14
Inferior	97.95±1.17	97.63±2.03
Inferio–lateral	96.67±0.34	99.56±0.13
Inferio–posterior	97.13±1.32	99.62±0.61
Inferio–posterior–lateral	96.33±3.5	93.37±6.16
Lateral	100±0	99.92±0.26
Posterior	99.56±0.73	99.38±0.5
Posterior–lateral	99.8±0.26	99.45±0.5
Healthy	99.95±0.04	99.80±0.06

The above–mentioned works represent application of the research and processing methods, which are specific for the big data during research. The data is taken from archives.

Recognition of Hypertension

To predict hypertension, Chinese researchers have used a variety of approaches, from which the random forest algorithm gave the best performance [13]. Information was collected from patients undergoing a stress test on a treadmill [14]. These data sets include verbal physical data, a description of the patient complaints (chest pain, shortness of breath, etc.), instrumental indicators during the test, and the anamnesis. The interquartile range was used to remove the anomalies. This method works well for "symmetric" data, in which the median is equal to the average value of the span. The gaps in the diastolic blood pressure values were replaced by the average values in the sampling. A discretization was also performed [15]. Six approaches were considered for the disease classification: the neural network with back–propagation errors, LogitBoost, the locally weighted naive Bayes, the Bayesian network, the support vector machine and the random forest methods. As an algorithm for the selection of signs, the algorithm for ranking by the information gain indicator [16] was chosen, and then–by their individual assessment [17]. The mean square error and the ROC curve were chosen as the result evaluation metric. As a result of the selection of signs, the most informative signs were identified (Figure 2).

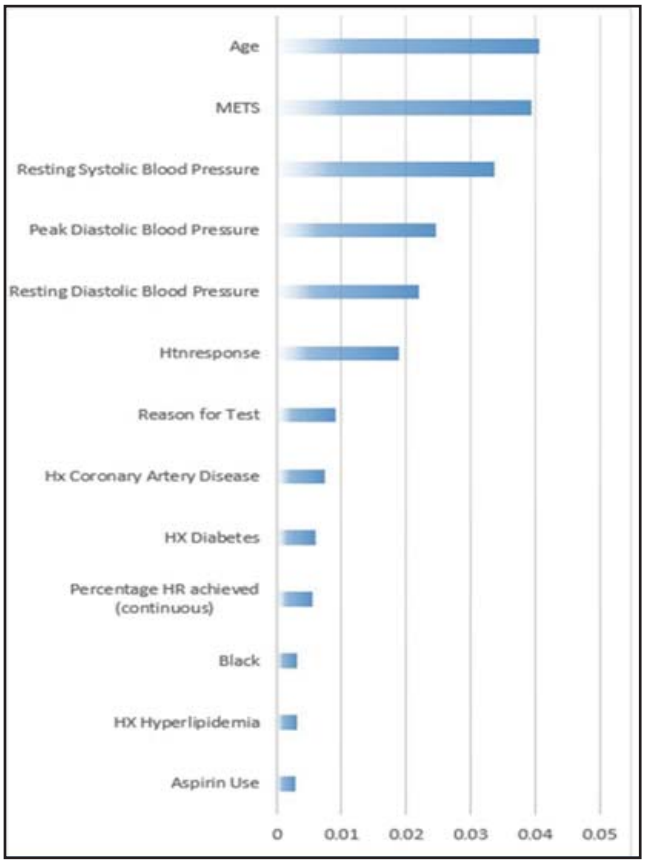

Figure 2: Ranking of signs result [13]

The random forest algorithm showed the best result (AUC=0.93). The results of all algorithms are presented in Table 4.

Table 4: Comparison of the various algorithms results

	ANN	LB	LWB	RTF	BN	SVM
Sensitivity	40%	31.3%	43%	75%	49.5%	28.2%
Specificity	88.4%	88.5%	80.92%	86.2%	79.8%	88.9%
Prediction	65.2%	59.3%	54.8%	73%	56.8%	57.7%
F–score	49.8%	40.9%	48.23%	74%	52.9%	37.9%
AUC	0.74	0.7	0.7	0.89	0.72	0.59
RMSE	0.44	0.45	0.46	0.46	0.42	0.57

The use of a large number of medical parameters is specific in this work, as well as the use of research methods specific for the big data. As a result, a comparative characteristic of a number of different classification algorithms was obtained for the diagnostic purpose.

Oxygen Consumption Dynamics

To reveal the dynamics of oxygen consumption by a human, a forecasting method based on the random forest algorithm was applied [18]. This technique will allow forecasting the deviation of human health from the norm at an early stage. The overall concept of the forecast system is shown in Figure 3.

To obtain data, the tests were conducted, during which the subjects, by excessive walking and other cardio–loadings, produced an imitation of daily activity. In this case, such indicators as heartbeat, minute volume of lung ventilation, respiration rate, hip dynamics and frequency of steps were obtained.

To evaluate the results, the MNG (mean normalized gain amplitude) metric was used. It was used as an indicator of the consumption dynamics too. The comparison results of the predicted dynamics and the real dynamics are shown in Figure 4.

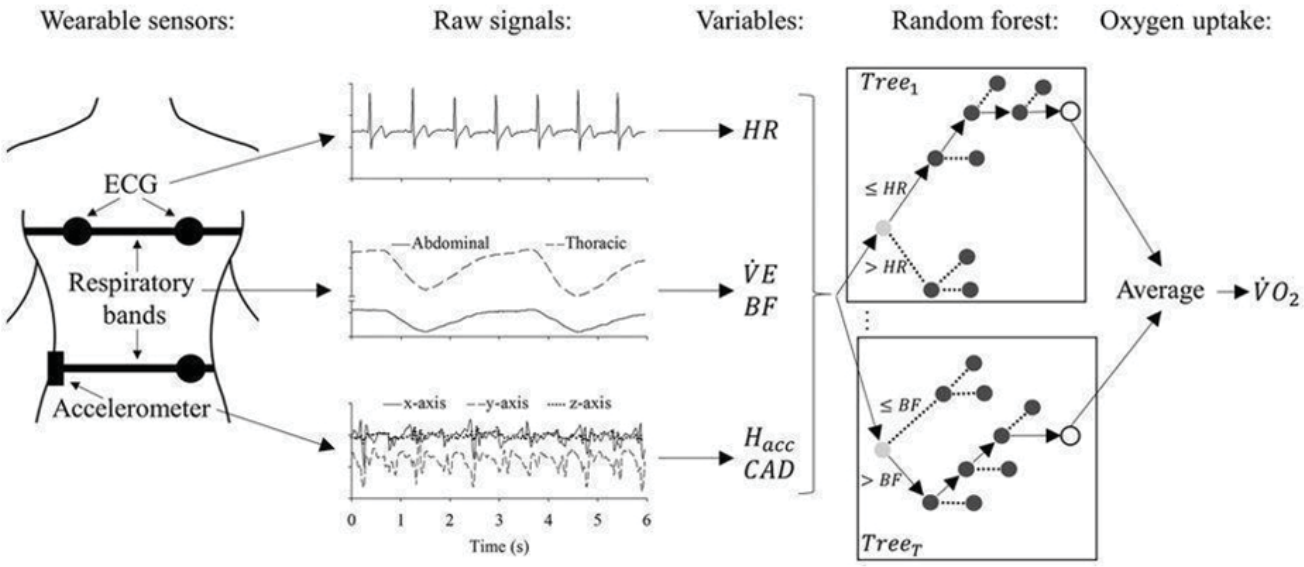

Figure 3: The concept of the oxygen consumption forecast system [13]

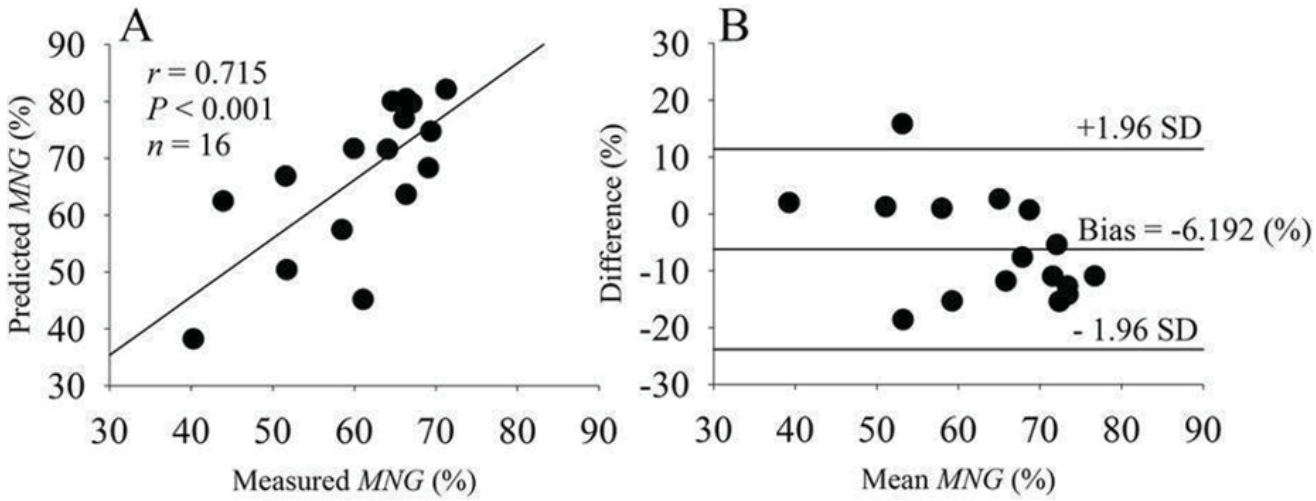

Figure 4: Correlation between the predicted MNG and the real MNG [18]

In this work, the data was obtained as a result of medical measurements using body sensors. Despite the relatively small amount of data analyzed, due to the applied technique, a practically significant result was obtained.

Fall Detection

Older people face the problem of falling, at home or in the street. In such situations, the issue arises of an adequate response to emergency situations of this kind for vital signs monitoring systems. For this purpose, a method was elaborated that allows not only to determine the fall condition, but also to classify falls by the nature. Moreover, this data will help therapists to understand reasons of the fall for better care [19].

The algorithm consists of 2 basic steps: at the first stage, smart textile technology allows to collect all the necessary data about a person (coordinates from accelerometer, breathing and heartbeat parameters). At the second stage, if the expected fall occurred, the support vector machine method classifies the fall, and assigns it to one of the classes (during ascent, during descent, during walk, during jogging, standing, falling forward, falling backwards, falling to the right, falling to the left, lying, sitting).

The result metrics are as follows: accuracy=98 %, sensitivity=97.6 %, specificity =98.5 %.

During this work the data was collected in the process of the research. The range of measurable indicators is insignificant. The key to getting results is the use of wearable electronics.

RECOGNITION OF THE HUMAN CONDITION

The big data processing methods can be used not only to forecast the occurrence or non–occurrence of a patient's disease, but also to predict future cognitive (psychological) conditions.

Detection of Cognitive Conditions

To identify cognitive conditions, the nearest neighbor and the random forest methods were used [20]. The Minko

vsky distance was chosen as the distance metric for the nearest neighbor method.

To collect data, the following approach was used: on the subjects were put sensors, which read vital signs (heart rate, respiration rate, frequency of steps, etc.). After receiving the data, the subjects recorded their condition through a special mobile application. Thus, a sampling for training was obtained.

The data for training and testing was divided for 10 blocks using cross–validation. The recognition results are shown in Table 5.

Table 5: Comparison of the various algorithms results

Patient	Classification method	AUC	Accuracy
A	The nearest neighbor method	0.6/0.0049	58.3/0.52
B		N/A	N/A
C		0.46/0.0067	56.5/0.67
D		0.55/0.0032	54.2/0.41
E		0.63/0.0034	59.7/0.40
A	The random forest method	0.65/0.0049	60.0/0.57
B		N/A	N/A
C		0.61/0.0080	56.2/0.78
D		0.70/0.0032	64.1/0.38
E		0.64/0.0037	60.4/0.44

The work also relies on the use of wearable electronics and processing techniques, special for the big data processing. Some other works are also dedicated to identify the patient condition. In particular, an electrocardiogram may be used to detect the emotional condition [21]. Such works are focused on getting results in an emergency mode.

PROSPECTS

As can be seen from the above review, in most studies only certain aspects of the big data technology are applied. These are usually mathematical processing methods, such as neural networks of various architecture and classification methods.

As a rule, algorithms are elaborated on the basis of archival data, which is examination data, in particular cardiograms. In the work dedicated to the study of the oxygen consumption dynamics and in the works on the recognition of the patient psychological condition, the data was obtained directly from the experiment. However, the duration of the experiment was not long.

Use of the data from various patients and short–term individual observations do not allow conducting studies aimed at forecasting the patient condition dynamics.
It should also be noted, that even small amount of data, both in terms of duration and in the number of parameters, collected by wearable devices, make it possible to obtain significant medical results.

We can say that the direction, associated with the data collection by wearable devices during long–term observations of the patient, is relevant and promising in the big data medicine. At the same time, it is obvious that the more measured parameters can be observed, the more significant results will be obtained.

METHODS

As can be seen from the above–listed medical data processing methods, the main methods are machine learning algorithms that allow to implement classification and prediction of various diseases.

Algorithms of artificial neural networks and, in particular, deep learning can also be used to solve problems of the disease recognition. Unlike machine learning, neural networks allow solving nontrivial tasks in conditions of noise or lack of data, but they, in general, sacrifice the accuracy of the results.

RESULTS

This article discusses the various applications of the big data technology in the field of medicine and health care associated with a system for monitoring and analyzing the patient medical data. Typical processing methods such as neural networks and machine learning have been identified.

DISCUSSION

The results obtained in this article show, that the topic of the research in the field of data collecting and processing of wearable biosensors using the big data technology is being actively developed and currently can be used in some areas of medicine to partially automate decision–making procedures. For each specific medical task it is possible to find the big data technology application, in most cases sufficient for acceptable decision–making, with appropriate accuracy and reliability.

ACKNOWLEDGEMENT

This research is performed with the financial support of the Ministry of Education and Science of Russia under the agreement No. 14.577.21.0232 dated September 29, 2016 (unique number RFMEFI57716X0232), applied scientific research is conducted on the topic "Research of scientific and technical solutions and elaboration of an intelligent biosensor platform for preventive monitoring and assessment of indicators of the human body "Body Sensor Network" with possibility to correlate data obtained from various sensors in a noisy environment."

NOMENCLATURE

Sensitivity. This metric shows the ratio of correctly recognized "unhealthy" patients to all "unhealthy" patients in the sampling;
Specificity. This metric shows the ratio of correctly recognized "healthy" patients to all "healthy" patients in the sampling;
Accuracy. This metric shows the ratio of correctly recognized patients to all patients in the sampling;
True positive (TP). The number of correctly recognized "unhealthy" patients;
True negative (TN). The number of correctly recognized "healthy" patients;
False positive (FP). The number of uncorrectly recognized "unhealthy" patients;
False negative (FN). The number of uncorrectly recognized "healthy" patients;
Precision. The ratio of correctly recognized "unhealthy" patients to all recognized as "unhealthy" patients;
Recall. The ratio of correctly recognized "unhealthy" patients to all "unhealthy" patients in the sampling;
Positive predictive value (PPV). The probability that a patient, recognized as "unhealthy", does have a disease;
F–score. The harmonic mean between accuracy and recall. The ratio of the accuracy and recall product, multiplied by 2 to the sum of accuracy and recall;
Area under the ROC curve (AUC). The correlation of the TP from the FP.

REFERENCES

1. Drowning in Big Data? Reducing Information Technology Complexities and Costs for Healthcare Organizations. (2012). Frost & Sullivan, from https://www.emc.com/collateral/analyst–reports/frost–sullivan–reducing–information–technology–complexities–ar.pdf, accessed on 2018-09-20.
2. Song, Z., Liu, C.H., Wu, J., Ma, J., Wang, W. (2014). QoI–aware multi–task–oriented dynamic participant selection with budget constraints. IEEE Transactions on Vehicular Technology, vol. 63, no. 9, 4618-4632, DOI: 10.1109/TVT.2014.2317701
3. Banaee, H., Ahmed, M.U., Loutfi, A. (2013). Data mining for wearable sensors in health monitoring systems: A review of recent trends and challenges. Sensors, vol. 13, no. 12, 17472-17500, DOI: 10.3390/s131217472
4. Rathi, M., Narasimhan, B. (2017). Data Mining, Soft Computing, Machine Learning and BioInspired Computing for Heart Disease Classification/Prediction – A Review, DOI: 10.23956/ijarcsse/V7I4/0156
5. Poon, C.C., Lo, B.P., Yuce, M.R., Alomainy, A., Hao, Y. (2015). Body sensor networks: In the era of big data and beyond. IEEE Reviews in Biomedical Engineering, vol. 8, 4-16, DOI: 10.1109/RBME.2015.2427254
6. Raghupathi, W., Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. Health Information Science and Systems, no. 2, 3, DOI: 10.1186/2047-2501-2-3
7. Yicuan, W., Terry, L., Terry, K., Byrd, A. (2015). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Technological Forecasting and Social Change, vol. 126, 3-13, DOI: https://doi.org/10.1016/j.techfore.2015.12.019
8. Zeinab Arabasadi, Roohallah Alizadehsani, Mohamad Roshanzamir, Hossein Moosaei. (2017). Computer aided decision making for heart disease detection using hybrid neural network–Genetic algorithm. Computer Methods and Programs in Biomedicine, vol. 141, 19-26, DOI: 10.1016/j.cmpb.2017.01.004
9. Blake, C., Merz, C. (1998). UCI Repository of Machine Learning Databases. University of California, Department of Information and Computer Science, Irvine.
10. Acharya, U.R., Fujita, H., Oh, S.L., Hagiwara, Y., Tan, J.H., Adam, M. (2017). Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals. Information Sciences, vol. 415-416, 190-198, DOI: https://doi.org/10.1016/j.ins.2017.06.027
11. Goldberger, A.L., Amaral, L.A.N., Glass, L., Hausdorff, J.M., Ivanov, P.C.H., Mark, R.G., Mietus, J.E., Moody, G.B., Peng, C.K., Stanley, H.E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation, vol. 101, no. 23, e215-e220, DOI: 10.1161/01.CIR.101.23.e215
12. Noorian, A., Dabanloo, N.J., Parvaneh, S. (2014). Detection and localization of myocardial infarction using K–nearest neighbor classifier. Conference "Computing in Cardiology 2014".
13. Sakr, S., Elshawi, R., Ahmed, A., Qureshi, W.T., Brawner, C., Keteyian, S., Blaha, M.J., Al-Mallah, M.H. (2017). Using machine learning on cardiorespiratory fitness data for predicting hypertension: The Henry Ford ExercIse Testing (FIT) Project. PLoS ONE, vol. 13, no. 4, e0195344, DOI: https://doi.org/10.1371/journal.pone.0195344
14. Al–Mallah, M.H., Keteyian, S.J., Brawner, C.A., Whelton, S., Blaha, M.J. (2014). Rationale and de

sign of the Henry Ford Exercise Testing Project (the FIT project). Clinical Cardiology, vol. 37, no. 8, 456-461, DOI: https://doi.org/10. 1002/clc.22302

15. Kurgan, L., Cios, K.J. (2001). Discretization algorithm that uses class-attribute interdependence maximization. Proceedings of the 2001 International Conference on Artificial Intelligence, p. 980-987.
16. Kent, J.T. (1983). Information gain and a general measure of correlation. Biometrika, vol. 70, no. 1, 163-173, DOI: https://doi.org/10.1093/biomet/70.1.163
17. Guyon, I., Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, no. 3, 1157-1182, DOI: 10.1162/153244303322753616
18. Beltrame, T., Amelard, R., Wong, A., Hughson, R.L. (2017). Prediction of oxygen uptake dynamics by machine learning analysis of wearable sensors during activities of daily living. Scientific Reports, no. 7, 45738, DOI: 10.1038/srep45738
19. Webster, E., Sukaviriya, N., Chang, H.-Y., Kozloski, J. (2017). Predicting cognitive states from wearable recordings of autonomic function. IBM Journal of Research and Development, vol. 61, no. 2/3, 2:1-2:11, DOI: 10.1147/JRD.2017.2648698
20. Apache Hadoop, from http://hadoop.apache.org/, accessed on 2018-09-21.
21. Chen, M., Ma, Y., Song, J., Lai, C.-F., Hu, B. (2016). Smart clothing: Connecting human with clouds and big data for sustainable health monitoring. Mobile Networks and Applications, vol. 21, no. 5, 825-845, DOI: https://doi.org/10.1007/s11036–016–0745–1

Paper submitted: 22.10.2018.
Paper accepted: 27.11.2018.