Biological Data Mining and Its Applications in Pulmonology

Oleksandr Komlevoi[a], Nataliia Komleva[b], Vira Liubchenko[b] and Svitlana Zinovatna[b]

[a] *Odesa National Medical University Valikhovsky Lane 2, Odesa, 65082, Ukraine*
[b] *Odesa Polytechnic State University, Shevchenko Ave., 1, Odesa, 65044, Ukraine*

Abstract

The processing of diagnostic data in pulmonology is complicated for a doctor due to the need to analyze many indicators, the relationships between which can be complex, and the degree of influence on the diagnostic result can be different. Traditionally, general clinical, biochemical, and questionnaire methods are used to support making a diagnosis. They allow describing the state of the bronchopulmonary system by a variety of indicators. The modern practice of laser correlation spectroscopy makes it possible to expand various indicators. Still, their values are represented by sets of one-dimensional distribution diagrams and are not convenient for analysis. Therefore, we investigated the feasibility of classifying the values of 32 biophysical indicators obtained by laser correlation spectroscopy in this work. We first performed data visualization and found that the classes of diagnosed diseases did not have a clear separation but were separated from the normal state. We then examined the results of classifying the data using three algorithms – naive Bayes, logistic regression, and random forest. We conclude that the most appropriate algorithm is logistic regression. The work value lies in expanding the set of diagnostic indicators due to the high-precision results of the classification of biophysical indicators, which increases the objectivity of the diagnosis of pulmonological diseases.

Keywords 1

Bioinformatics, pulmonological diagnostics, data analysis, biophysical indicators, classification

1. Introduction

"Bioinformatics is the combination of health information, data, and knowledge. It uses computational techniques and tools to analyze the enormous biological databases"[1]. Health care analysts aim to provide the best medical care at an affordable cost, using a modern material and technical base.

In [2], three goals of bioinformatics are listed: 1) organize data in such a way that "allows researchers to access existing information and to submit new entries as they are produced"; 2) "to develop tools and resources that aid in the analysis of data"; 3) "to use these tools to analyze the data and interpret the results in a biologically meaningful manner."

According to [3], "biological research involves the application of some type of mathematical, statistical, or computational tools to help synthesize recorded data and integrate various types of information in the process of answering a particular biological question."

Bioinformatics is used in various fields of biology. Its application in genomics is described in [4], in the field of plant breeding – in [5], in the development of drugs – in [6]. In [7], an overview of many areas of human activity in which bioinformatics can be applied is given. Biologists are stepping up their efforts in understanding the biological processes that underlie disease pathways in clinical contexts [8].

IDDM-2021: 4th International Conference on Informatics & Data-Driven Medicine, November 19–21, 2021, Valencia, Spain
EMAIL: shurik73.jan@gmail.com (A. 1); nkomlevaya@gmail.com (A. 2); lvv@op.edu.ua (A. 3); zinovatnaya.svetlana@op.edu.ua (A. 4)
ORCID: 0000-0002-8297-089X (A. 1); 0000-0001-9627-8530 (A. 2); 0000-0002-4611-7832 (A. 3); 0000-0002-9190-6486 (A. 4)


Recent advances in bioinformatics and computational biology have accelerated the development of complex biomedical systems (for example, knowledge-based decision systems and artificial intelligence in medical informatics). Artificial Intelligence (AI) has a significant impact on biological and medical research, and AI-based methods have been increasingly implemented in real-life medical/healthcare applications. [9] presents five innovative scientific papers related to artificial intelligence methodologies, with potential applications in biological and medical fields.

[10] also shows bioinformatics approaches for analyzing measurable indicators using mathematical, network solutions, and machine learning theories.

The state of the respiratory system is assessed by a set of biophysical indicators - the percentage of particles of various natures present in the exhaled air. They are obtained using the method of laser correlation spectroscopy. For a differential assessment of the physiological and pathological states of the respiratory system, the change in the biophysical parameters of the humidity of the exhaled air is assessed.

The work aims to improve the results of the analysis of the biophysical parameters of exhaled air condensate obtained by laser correlation spectroscopy in pulmonological diagnostics through intelligent data processing methods.

In general, a pulmonological diagnostic is an extensive area. Several works consider the principles of constructing machine learning algorithms to diagnose and predict asthma [11] and bronchitis [12]. In this article, a diagnostic analysis of the biophysical parameters of exhaled air condensate is carried out in the presence of these diseases.

2. State of the Art in Pulmonology

Currently, asthma and bronchitis research uses protocols that include the following methods [13, 14]:

- general clinical methods: analysis of data obtained from case histories on complaints, anamnesis of life and disease, objective examination, X-ray examinations;
- questionnaires: conducted with the help of universal questionnaires developed using socio-demographic methods;
- biochemical with the study of indicators of lipid peroxidation system (the content of primary products of lipid peroxidation – diene conjugates, secondary products – malonic dialdehyde), and the activity of the enzyme antioxidant protection – catalase.

For a wide range of respiratory diseases, including reversible and irreversible obstructive diseases, there are many studies on the pathogenesis, disorders of respiratory mechanics, ventilation-perfusion relations, respiratory function, and ventilation regulation.

Diagnosis of diseases and assessment of their severity is based on various clinical data and additional studies: a functional examination of the lungs, chest radiography, blood culture, computed tomography, serological and cultural studies, analysis of sputum by Gram, acid-base composition, and saturation level, the study of arterial blood gas composition, bronchoscopy, invasive procedures to obtain diagnostic material, routine laboratory tests, and other [15, 16].

Traditional methods for assessing the activity of inflammation in the airways include analysis of normal and induced sputum, bronchoalveolar lavage, bacterioscopy, and bronchobiopsy [17]. Some physicochemical methods allow to determine traces of gaseous substances in human exhaled air and determine its micro composition: mass spectrometry combined with gas chromatographic separation, gas chromatography, electrochemical sensors, UV chemiluminescence, and IR spectroscopy [18]. The latter includes Fourier spectroscopy, optoacoustic spectroscopy, and laser correlation spectroscopy, one of the most sensitive methods to study the composition of polydisperse liquids.

Until now, laser correlation spectroscopy has been used to study the molecular composition and functional characteristics of various biological fluids (blood plasma, oropharyngeal washes, etc.) [19]. However, to assess the state of the human pulmonological system, it would be advantageous to study the moisture condensate of the air exhaled by a person using laser correlation spectroscopy.

In earlier works, attempts were made to study the state of the pulmonological system based on the analysis of exhaled breath condensate subfractions. However, the conclusions were based on a rather

small number of indicators (from 3 to 6), which led to about 15% of errors of the second kind nd a classification precision of about 85% [20, 21].

3. Research methods and models

Condensation of exhaled air moisture is formed by condensation of water vapor from the environment and aerosol impurities, which appear when the liquid is adjacent to the epithelial layer of the respiratory tract. The primary source of this aerosol impurity is the distal part of the respiratory system. Therefore, changes in the condensate components reflect changes in the level of small bronchi and alveoli.

The general principles of the respiratory system condition analysis of the inspected person consist of the following. Previously, the study of a representative sample of persons without comorbidities (conditional norm) establishes the most common nature of the relationship of different sizes of biological ingredients, which is normological. All other different characteristics of the relationship of ingredients reflect a variety of changes.

Biophysical indicators are the percentage contributions of microparticles entering the moisture condensate of the exhaled air. Their redistribution means the presence or absence of various macromolecular fractions, which corresponds to different states of the bronchopulmonary system, including the presence of pathologies.

To obtain numerical values of biophysical parameters, microparticles of exhaled air were examined by laser correlation spectroscopy. The technique of laser correlation spectroscopy is based on the measurement of the spectral characteristics of monochromatic coherent radiation due to the scattering of light as it passes through a dispersed nanoparticle system suspended in a liquid.

By measuring the fluctuation spectrum of the photocurrent and determining its half-width, it is easy to obtain the particle size in the system under study. As a rule, the investigated samples are polydisperse, i.e. at the same time in a solution there are particles of different sizes. However, the studied particle sizes (especially biological fluids) are rarely monodisperse. Suppose the spectrum of light scattered by monodisperse particles is a Lorentz curve (I), then for a polydisperse system. In that case, the spectrum is a sum, and for continuous distributions – an integral of Lorentzians with different half-widths G. In this case, the spectrum has the form:

$$I(W) = \int \frac{A(\tilde{A})\tilde{A}}{\tilde{A}^2+\omega}\, d\tilde{A}, \quad (1)$$

where $A(\tilde{A})$ – particle distribution function by diffusion coefficients and, consequently, by size. Determining the size distribution of particles is thus the solution of this integral equation with the Lorentz nucleus.

A self-developed device is used to collect condensate of exhaled air moisture, which allows obtaining the same volumes of condensate in a short time from each of the subjects in the amount required for the laser correlation spectroscopy method [22].

The moisture condensate of the exhaled air, like other biological fluids, is rarely monodisperse. Usually, polydisperse samples are examined, i.e., particles of different sizes are simultaneously in solution. The size distribution of condensate moisture in exhaled air particles is a histogram defined on a grid of 32 points. The high accuracy of the method allows showing the concentration of particles with a radius of 1 to 10000 nm.

The obtained biophysical parameters formed the basis for diagnosing the condition of the pulmonological system of the subjects. Diagnosis is performed by classifying the conditions of the subjects, using three classes: “norm,” “asthma,” and “bronchitis.”

Traditionally, much biological and medical research is based on Naive Bayes Classifier [23]. It allows predicting posterior probabilities of specific class membership [24, 25].

No less common is the use of logistic regression in diagnostics [26, 27]. “A major advantage of logistic regression compared to other similar approaches like probit regression—and therefore, a reason for its popularity among medical researchers – is that the exponentiated logistic regression slope coefficient can be conveniently interpreted as an odds ratio. The odds ratio indicates how much the odds of a particular outcome change for a 1-unit increase in the independent variable” [28]. In [29], it is said

that the need for preliminary data analysis and it is shown that the excessive complexity of the dataset leads to low classification accuracy.

Using a classifier based on a random forest is successful, as it demonstrates its effectiveness when processing data with a large number of signs [30]. In addition, in the future, random forest methods can make it possible to assess the significance of individual characteristics and reduce their number while ensuring the required classification accuracy [31]. As shown in [32], the use of random forest will improve the explainability of the results for users who are not experts in the studied subject area.

4. Experiment and results

In studies of asthma and bronchitis diseases using laser correlation spectroscopy, 32 biophysical parameters can be monitored. However, the approach used makes it possible to analyze the situation for each indicator separately, which can lead to a distortion of the diagnosis due to ignoring the influence of other indicators. Accordingly, there is a need to automate the classification procedure to ensure the possibility of simultaneous accounting of all available indicators.

The experiment performed consisted of two stages. In the first step, we visualized the available data to determine how different the classes under study differ. In the second step, we classified the available data using three algorithms to assess the best classification results.

4.1. Preliminary analysis of pulmonary data

Under the study's objectives, the entire contingent of the surveyed was divided into three non-overlapping groups: children aged 6 to 10 years whose parents agreed to the processing of pulmonary data. The first group included 275 children who were not burdened with known diseases based on a pediatric examination. The second group consisted of 115 children who had bronchial asthma. The third group included 131 children with bronchitis.

The quality of the samples obtained was assessed by two indicators: representativeness and reliability. A serial sampling methodology was followed to ensure representativeness, which allows reducing the determinism in sampling and approaching random sampling. At the same time, objectively existing groups of healthy children and children with asthma and bronchitis are randomly selected. All children were examined within the groups. To ensure reliability:

- ensured that all elements of the population were included in the sampling frame;
- checks were carried out for the absence of duplication of the surveyed;
- guaranteed exclusion from the sampling frame of the surveyed groups that did not meet the requirements (children with unconfirmed asthma and bronchitis or unconfirmed healthy conditions);
- created and centrally stored electronic lists with unambiguous identification of the surveyed data.

Let us investigate the ability of biophysical traits to separate classes of diseases. To visually represent the values of 32 biophysical traits for different classes of diseases, we will reduce the dimension of the initial data while minimizing the loss of information. For this, we use the principal component analysis (PCA) method. Figure 1 shows the results of class division using two components based on a representative sample of initial data. The values of biophysical parameters ideally identify the class "norm" for the classes "asthma" and "bronchitis," there is some overlap. It means that the values of biophysical parameters can be used to diagnose asthma and bronchitis.

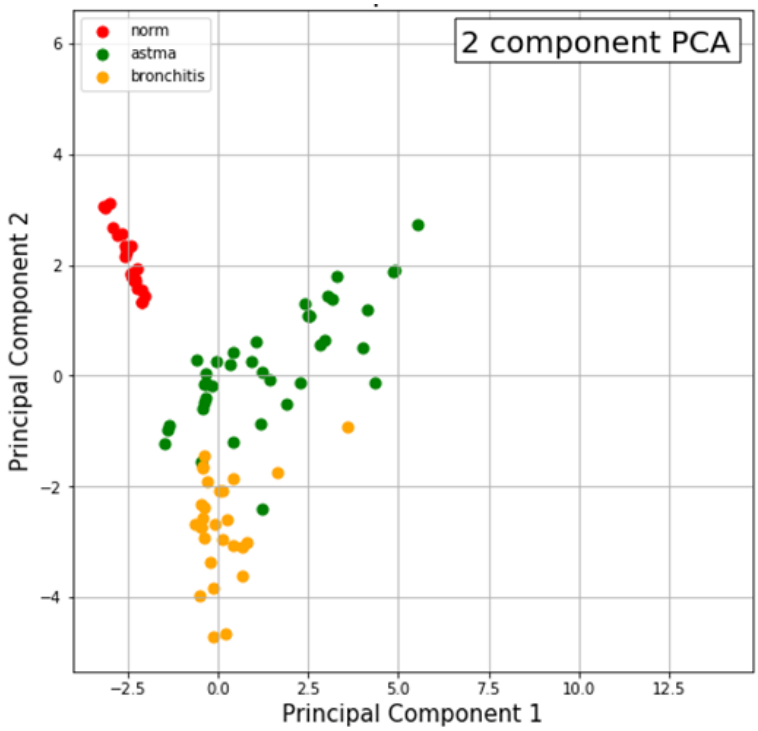

Figure 1: Principal component distribution of classes data

4.2. Pulmonary data classification

Cross-validation is used to evaluate the model's training. The original data is divided into subsets Train, Valid, and Test (Figure 2).

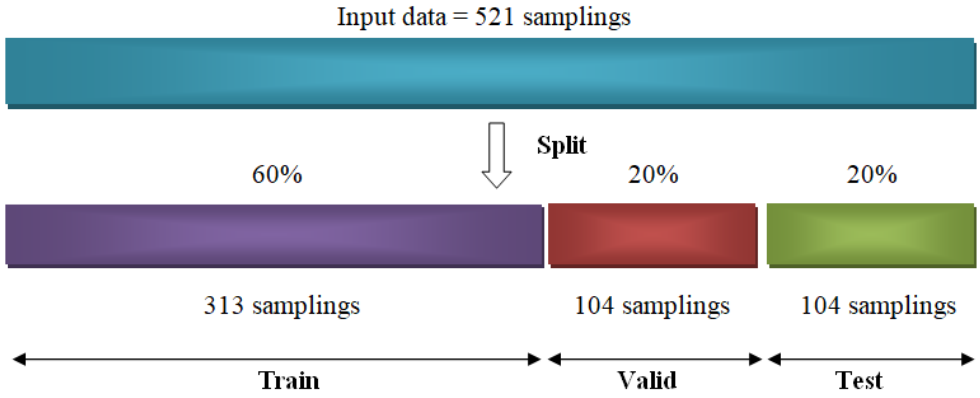

Figure 2: Distribution of data in constructing classifiers

Table 1 shows the characteristics of the constructed classifiers: values of accuracy, completeness, and F1-measure as a harmonic mean between them.

Table 1
Primary characteristics of classifiers

Classifier	Precision	Recall	F1-measure
Naive Bayes	0.783	0.761	0.772
Logistic Regression	0.870	0.823	0.846
Random Forest	0.859	0.844	0.851

As expected, the Naive Bayes Classifier performed the worst because of a natural interdependence between the values of various biophysical signs. However, these dependencies have not yet been formalized.

In order to reduce errors of classifiers, two approaches were used:

- putting forward more stringent requirements for experimental conditions;
- balancing the studied classes.

Since the experimental study of the moisture condensation of exhaled air was carried out in various medical institutions and using different (albeit of the same type) equipment, it was decided to use only those data that were obtained on equipment with completely identical characteristics. In addition, for the "asthma" and "bronchitis" classes, patients who at the time of the examination had just been admitted to the hospital and had not yet begun treatment were considered.

To take into account all these factors, a database was developed, the diagram of which is shown in Figure 3.

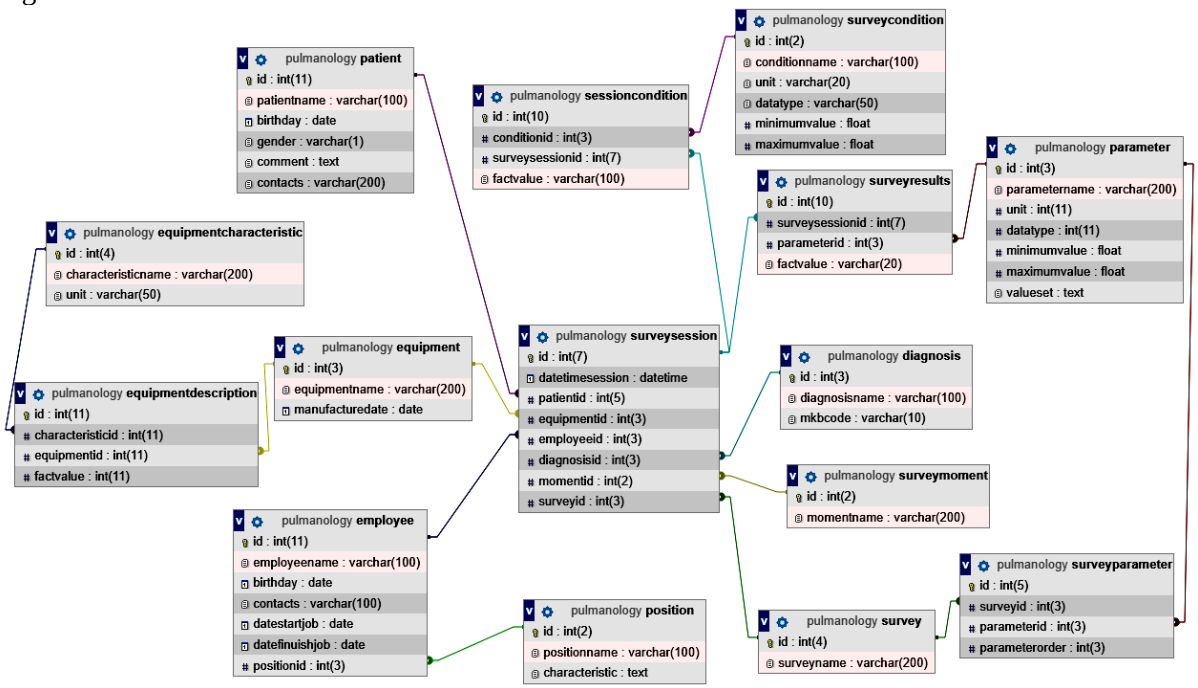

Figure 3: The patient survey database schema

The database includes reference tables. For example, the table «surveymoment» contains characteristics of the survey time in relation to the patient's state: upon admission to the hospital, immediately after treatment, after two weeks for adaptation after the treatment end, etc. The characteristics of the survey conditions include temperature, relative humidity, time of patient adaptation before the examination, etc. The values of the characteristics of the survey conditions have different measurement units and can have the specified limits of the value range.

"Sifting" the initial experimental data made it possible to improve the quality of the classification; the obtained modified characteristics of the classifiers are shown in Table 2.

Table 2

Modified characteristics of classifiers

Classifier	Precision	Recall	F1-measure
Naive Bayes	0.814	0.796	0.805
Logistic Regression	0.933	0.918	0.925
Random Forest	0.874	0.862	0.868

The studied classes are not balanced, affecting the quality of the solution to the classification problem. There are several approaches to overcome the imbalance problem, e.g., undersampling

(reduction of large classes) and oversampling (duplication of objects of small classes) to equalize the proportions of the studied classes.

The use of the undersampling approach leads to a decrease in the number of samples for the "norm" class by more than 2 times, and for the class "bronchitis" – approximately 1.2 times to match the dimension of the minor class "asthma." The oversampling approach can cause an overfitting problem if duplicates of minor class objects fall into the Train, Valid, and Test sets. Therefore, the Synthetic Minority Oversampling Technique (SMOTE) is an effective method, which interpolates existing objects of the minor class and expands their size at the expense of the resulting objects. The results of applying the undersampling and oversampling approaches are shown in Figure 4.

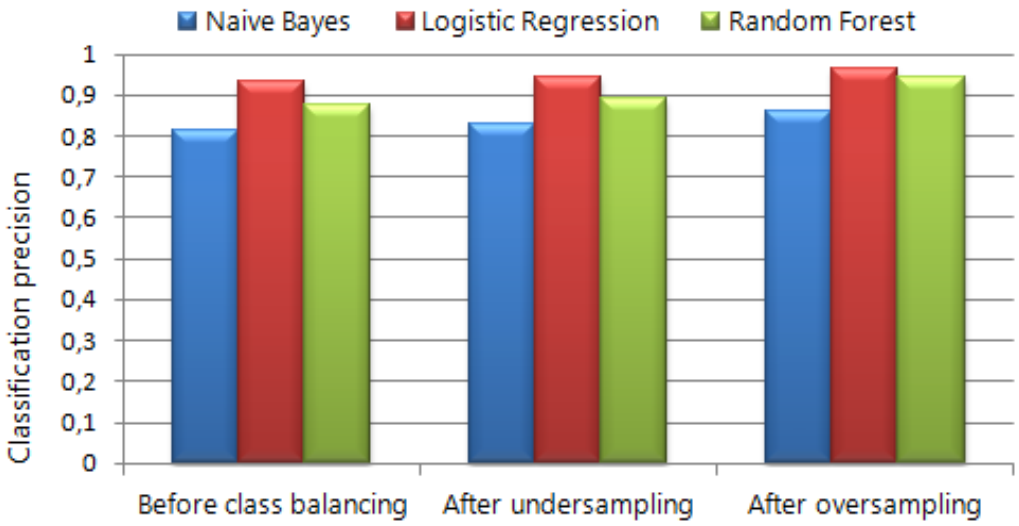

Figure 4: Classification precision values for classifiers of different architectures

The precision on all classifiers increased; the most significant increase in precision was obtained after applying the SMOTE method.

The received results were validated with the multi-class cross-entropy loss calculated as:

$$L(X_i, Y_i) = -\sum_{j=1}^{c} y_{ij} * log(p_{ij}), \quad (2)$$

where Y_i is a one-hot encoded vector ($y_{i1}, y_{i2}, \ldots, y_{ic}$),

$y_{ij} = \begin{cases} 1 & \text{if } i\text{th element is in class } j \\ 0 & \text{otherwise} \end{cases}$,

p_{ij} is the probability that ith element is in class j.

The validation confirmed the results described above.

5. Discussion

Making a diagnosis is extremely important and risky since the patient's health, and sometimes life depends on its solution's correctness. Therefore, the informational support of the doctor in the process of making a diagnosis is valuable.

Improvement of pulmonological diagnostics means and the transition to laser correlation spectroscopy in studies of asthma and bronchitis diseases increased the number of indicators available for analysis by 32 biophysical indicators.

To improve the classification accuracy, the following procedures were carried out:

- selection of the classifier providing the best values of Precision, Recall and F1-measure;
- "sifting" the obtained experimental data;
- balancing classes.

When choosing a classifier, the best results were shown by logistic regression, which provided precision=0.870, recall=0.823, and F1-measure=0.846.

When "sifting" the data, the results of the following surveys were excluded:

- obtained on equipment with characteristics that differ from standard ones (laser power, centrifuge speed, etc.);
- obtained under incomplete observance of the experimental conditions (temperature, relative humidity of the room, etc.);
- obtained during the examination of patients who have already started treatment or have undergone a course of treatment (for class "asthma").

When balancing classes, the best results were shown by oversampling approach with applying the SMOTE method. This made it possible to improve the classification precision up to 96%.

Let us analyze possible errors in the operation of the obtained classifiers. Regardless of the chosen method (Naive Bayes, Logistic Regression, or Random Forest), errors of the first and second kind are present during the classification.

Errors of the first kind correspond to situations in which the examination showed the presence of a disease, although the person is healthy (regarding diseases such as asthma and bronchitis). Since the proposed classifiers do not replace the doctor's decision but provide him with additional helpful information to help make a diagnosis, additional examinations and analyses can reveal errors of the first kind ("false alarm").

Errors of the second kind correspond to "missing the target" and are traditionally much more severe. In this case, the sick patient is diagnosed as healthy, and with the screening form of examination, he may not be provided with primary treatment.

The estimation of errors of the second kind when using the classifier based on logistic regression showed 7% errors before the procedures to improve the accuracy and 4% after these procedures, which is a good result.

6. Conclusion

The article is the first to apply the method of laser correlation spectroscopy, which allows studying the moisture condensate of the exhaled air by the values of 32 biophysical indicators. Based on the values of these indicators, the diagnosis of diseases for asthma and bronchitis in children aged 6 to 10 years was performed.

The traditional approach to analyzing indicators is based on determining the indicators with the most excellent separating ability and comparing the patient's data for these indicators with the "reference" ones. However, this approach has many disadvantages. Firstly, with the traditional approach, the comparison is carried out for 3-6 separate indicators. Secondly, whether the separating ability of individual indicators depends on the age of the patients and the region in which they live has not been investigated today. Therefore, most of the indicators remain unused in the diagnostic process. At the same time, it is known that the quality of a solution is determined mainly by the amount of information used to obtain it.

The article examined the methods of Naive Bayes, Logistic Regression, and Random Forest as applied to the diagnosis of diseases for asthma and bronchitis. Studies have shown that the classifier based on logistic regression showed the highest precision in classifying – 96%, despite previous studies giving up to 85%. At the same time, the number of errors of the second kind was reduced to 4%, i.e., they are reduced by more than three times. The use of classification methods for solving pulmonological diagnostics made it possible to consider all the available data obtained by laser correlation spectroscopy.

Studies have shown that using classification methods for biophysical indicators for diagnosing pulmonary diseases allows obtaining results with high accuracy, which improves the quality of diagnosis. Automation of the classification process provides fast, objective results.

Data visualization showed that the classes of diseases do not have a pronounced boundary. Still, they are clearly separated from the normal state. Therefore, a further direction of this work is to study the applicability of fuzzy classification methods for solving pulmonological diagnostics.

7. References

[1] V. Majhi, S. Paul, R. Jain, Bioinformatics for Healthcare Applications, in: Proceedings of the 2019 Amity International Conference on Artificial Intelligence (AICAI), IEEE 2019, pp. 204–207. doi:10.1109/AICAI.2019.8701277.

[2] N. M. Luscombe, D. Greenbaum, M. Gerstein, What is bioinformatics? An introduction and overview, Yearbook of Medical Informatics 10.01 (2001) 83–100. doi:10.1055/s-0038-1638103.

[3] J. Xiong, Essential Bioinformatics, Cambridge University Press, 2006.

[4] K. Gobalan, J. Ahamed. Applications of Bioinformatics in Genomics and Proteomics, Journal of Advanced Applied Scientific Research 1.3 (2016) 29–42.

[5] D. F. Gomez-Casati, M. V. Busi, J. Barchiesi, D. A. Peralta, N. Hedin, V. Bhadauria, Applications of Bioinformatics to Plant Biotechnology, Current Issues in Molecular Biology 27 (2018) 89–104. doi:10.21775/cimb.027.089.

[6] S. K. Gill, A. F. Christopher, V. Gupta, P. Bansal, Emerging role of bioinformatics tools and software in evolution of clinical research, Perspectives in Clinical Research 7.3 (2016) 115–122. doi:10.4103/2229-3485.184782.

[7] M. Younus Wani, NA. Ganie, S. Rani, S. Mehraj, M. R. Mir, M. F. Baqual, K. A. Sahaf, F. A. Malik, K. A. Dar, Advances and applications of Bioinformatics in various fields of life, International Journal of Fauna and Biological Studies 5.2 (2018) 03–10.

[8] F. Wang, X. Li, J. T. L. Wang and S. Ng, Guest Editorial: Special Section on Biological Data Mining and Its Applications in Healthcare, IEEE/ACM Transactions on Computational Biology and Bioinformatics 14.3 (2017) 501–502. doi:10.1109/TCBB.2016.2612558.

[9] H. Alinejad-Rokny, E. Sadroddiny, V. Scaria, Machine learning and data mining techniques for medical complex data analysis, Neurocomputing 276 (2018). doi:10.1016/j.neucom.2017.09.027.

[10] Y. Lin, F. Qian, L. Shen, F. Chen, J. Chen, B. Shen, Computer-aided biomarker discovery for precision medicine: data resources, models and applications, Briefings in Bioinformatics 5.3 (2019) 952–975. doi:10.1093/bib/bbx158.

[11] J. Finkelstein, I. C. Jeong, Machine learning approaches to personalize early prediction of asthma exacerbations, in: Annals of the New York Academy of Sciences 1387.1 (2017) 153–165. doi:10.1111/nyas.13218.

[12] H. S. Porieva, K. O. Ivanko, C. I. Semkiv, V. I. Vaityshyn, Investigation of Lung Sounds Features for Detection of Bronchitis and COPD Using Machine Learning Methods, Visnyk NTUU KPI Seriia-Radiotekhnika Radioaparatobuduvannia 84 (2021) 78–87.

[13] E. C. Oelsner, L. R. Loehr, A. G. Henderson, K. M. Donohue, P. L. Enright, R. Kalhan, C. M. Lo Cascio, A. Ries, N. Shah, B. M. Smith, W. D. Rosamond, Classifying Chronic Lower Respiratory Disease Events in Epidemiologic Cohort Studies, Annals of the American Thoracic Society 13.7 (2016) 1057–1066. doi:10.1513/AnnalsATS.201601-063OC.

[14] A. Kantar, Phenotypic presentation of chronic cough in children, Journal of Thoracic Disease 9.4 (2017) 907–913. doi:10.21037/jtd.2017.03.53.

[15] M. Pokorski, Chest Radiography in Children Hospitalized with Bronchiolitis, Pulmonology 1222 (2019) 55–62. doi:10.1007/5584_2019_435.

[16] M. Shafiq, H. Lee, L. Yarmus, D. Feller-Kopman, Recent Advances in Interventional Pulmonology, Annals of the American Thoracic Society 16.7 (2019) 786–796. doi:10.1513/AnnalsATS.201901-044CME.

[17] M. Heching, D. Rosengarten, D. Shitenberg, O. Shtraichman, N. Abdel-Rahman, A. Unterman, M. R. Kramer, Bronchoscopy for Chronic Unexplained Cough Use of Biopsies and Cultures Increase Diagnostic Yield, Journal of Bronchology & Interventional Pulmonology 27.1 (2020) 30–35. doi:10.1097/LBR.0000000000000629.

[18] P. Shende, J. Vaidya, Y.A. Kulkarni, R.S. Gaud, Systematic approaches for biodiagnostics using exhaled air, Journal of Controlled Release 268 (2017) 282–295. doi:10.1016/j.jconrel.2017.10.035.

[19] E. Nepomnyashchaya, E. Velichko, E. Aksenov, T. Bogomaz, Laser Correlation Spectroscopy as a Powerful Tool to Study Immune Responses, in: Proceedings Volume 10685, Biophotonics:

Photonic Solutions for Better Health Care VI, SPIE Photonics Europe, 2018, Strasbourg, France, 2018. doi:10.1117/12.2307241.
[20] N. Komlevaya, A. Komlevoy, K. Chernega, Designing of the specialized computer system for making pulmonology diagnosis, in: Proceedings of the 8th International Conference of Programming UkrPROG'2014, Kyiv Ukraine, 2014, pp. 253–263.
[21] N. O. Komleva, K. S. Cherneha, B. I. Tymchenko, O. M. Komlevoy, Intellectual Approach Application for Pulmonary Diagnosis, IEEE First International Conference «Data Stream Mining & Processing», Lviv Ukraine, 2016, pp. 48–52.
[22] O. M. Komlevoi, Yu. I. Bazhora, inventors; Odessa State Medical University, patent owner. Device for collecting moisture condensate from exhaled air. Ukraine patent UA № 47117. Filed November 6th., 2009, Issued January 1st., 2010.
[23] N. Boyko, K. Boksho, Application of the Naive Bayesian Classifier in Work on Sentimental Analysis of Medical Data, in: Proceedings of the 3rd International Conference on Informatics & Data-Driven Medicine, Växjö Sweden, 2020, pp. 230–239.
[24] S. R. B. Shree, H. S. Sheshadri, Diagnosis of Alzheimer's disease using Naive Bayesian Classifier, Neural Computing & Applications 29.1 (2018) 123–132. doi:10.1007/s00521-016-2416-3.
[25] B. Bratic, V. Kurbalija, M. Ivanovic, I. Oder, Z. Bosnic, Machine Learning for Predicting Cognitive Diseases: Methods, Data Sources and Risk Factors, Journal of Medical Systems 42.12 (2018). doi:10.1007/s10916-018-1071-x.
[26] E. Y. Boateng, D. A. Abaye, A Review of the Logistic Regression Model with Emphasis on Medical Research, Journal of Data Analysis and Information Processing 7 (2019) 190–207. doi:10.4236/jdaip.2019.74012.
[27] N. Esener, A. M. Guerra, K. Giebel, D. Lea, M. J. Green, A. J. Bradley, T. Dottorini, Mass spectrometry and machine learning for the accurate diagnosis of benzylpenicillin and multidrug resistance of Staphylococcus aureus in bovine mastitis, PLOS Computational Biology 17.6 (2021). doi:10.1371/journal.pcbi.1009108.
[28] P. Schober P, T. R. Vetter, Logistic Regression in Medical Research, Anesthesia & Analgesia 132.2 (2021) 365–366. doi:10.1213/ANE.0000000000005247.
[29] S. Kaur, A. Abdullah, N. N. Hairi, S. K. Sivanesan, Logistic Regression Modeling to Predict Sarcopenia Frailty among Aging Adults, International Journal of Advanced Computer Science and Applications 12.8 (2021) 497–504.
[30] Z. Khan, A. Gul, A. Perperoglou, M. Miftahuddin, O. Mahmoud, W. Adler, B. Lausen, Ensemble of optimal trees, random forest and random projection ensemble classification, Advances in Data Analysis and Classification 14.1 (2020) 97–116. doi:10.1007/s11634-019-00364-9.
[31] Y. Tian, J. Yang, M. Lan, T. Zou, Construction and analysis of a joint diagnosis model of random forest and artificial neural network for heart failure, AGING-US 12.24 (2020) 26221–26235. doi:10.18632/aging.202405.
[32] D. Petkovic, R. Altman, M. Wong, A. Vigil, Improving the explainability of Random Forest classifier – user centered approach, Pacific Symposium on Biocomputing 23 (2018) 204–215.