Cytological Image Classification Using Data Reduction

Oleh Berezsky[1[0000-0001-9931-4154]], Oleh Pitsun[2[0000-0003-0280-8786]], Taras Dolynyuk[3[0000-0003-1758-1203]], Lesia Dubchak[4[0000-0001-5763-9540]], Nadiya Savka[5[0000-0003-4182-7867]], Grygoriy Melnyk[5[0000-0003-0646-7448]], Vasyl Tesluk[7[0000-0002-5974-9310]]

[1,2,3,4,5,6] Ternopil National Economic University, 46001, Ukraine
[7] Lviv Polytechnic National University, 79000, Ukraine

ob@tneu.edu.ua
o.pitsun@tneu.edu.ua
trsdln@gmail.com
dlo@tneu.edu.ua
n.savka@tneu.edu.ua
mgm@tneu.edu.ua
vasyl.teslyuk@gmail.com

Abstract. In this paper, the authors investigate data reduction techniques using cytological image data for the purpose of further classification applying modern approaches. Cytological images are widely used in diagnosing cancerous and precancerous conditions of the breast. Classification of the whole image is a rather time consuming process, so the authors apply an approach when quantitative characteristics of micro-objects (cell nuclei) are used for classification. The authors carried out a comparative analysis of the classification of the cytological images based on the quantitative characteristics of their nuclei using modern classifiers. The main criteria for describing micro-objects (cell nuclei) are the following ones: area, perimeter, circumference, maximum width and length, area and perimeter of the bounding box. The structure of the biomedical image classification system is developed, including image processing stages, calculations of quantitative characteristics of micro-objects, data reduction and classification. The principal component method is used as data reduction technique. To classify data the following methods are used: a single-layer perceptron, logistic regression, support vectors machine, and the k-nearest neighbor method. Testing was performed using BPCI2100 database of cytological images of cancerous and precancerous conditions of the breast.

Keywords: Classification, Principal Component Method , Cytology, breast precancerous conditions.

1 Introduction

Cytological and histological images are used to diagnose precancerous and cancerous conditions of the breast. After a microscopic examination, a specialist can determine the type of an image. To simplify the analysis process of cytological and histological images, a number of automated microscopy systems (AMSs) with functions for image

2019 IDDM Workshops.

processing were developed. The use of artificial intelligence, in particular artificial neural networks, support vector machines, etc. show the current trends of improvement in AMSs [1]. This makes the process less time-consuming and allows us to increase the diagnostic efficiency. The main indicators of the pathology in cytological specimens are the shape and structure of the cells. The following quantitative characteristics of the investigated micro-objects such as area, perimeter, circumference, angle of inclination of the main axis, etc. are used. To calculate them, firstly, the input image is preprocessed (filtering, histogram alignment). In next stage, the following segmentation methods are used: threshold segmentation, watershed method, k-means method or their combinations. When the cytological image is converted into a binary format, each micro-object is detected and the quantitative characteristics are calculated. The training sample has a set of features. Array features include redundant and uninformative features. Therefore, more time is required for classification. So, it is necessary to reduce the input feature set. The following methods are used for feature reduction: complete search, depth-first search, breadth-first search, branch-and-bound, group method of data handling, feature ranking, feature clustering, evolutionary search, etc. [2]. In this paper, the principal component method is used for reduction of input characteristics. The following methods have been selected as high-level computer vision tools: support vector method, logistic regression, a single-layer perceptron, and the k-nearest neighbor method. Support vector method is a method of analyzing data for classification using directed learning models. Each element of the training samples is assigned to a certain class. The training algorithm creates a model that assigns new samples to one of the classes. Formally, the support vector machine builds a hyperplane, or a set of hyperplanes in high-dimensional space that can be used for classification, regression and other tasks [3]. Logistic regression is a statistical regression method used when a dependent variable is categorical, that is, it can have only two values [4]. The idea of logistic regression is that the space of the original values can be divided by a line into two corresponding classes. The k-nearest neighbor method assigns objects to the class that most of its k-nearest neighbors belong to in a multidimensional feature space. This is one of the simplest algorithms for learning classification models [5]. A single-layer perceptron is the simplest kind of artificial neural networks, which is based on a mathematical model of the information perception by the human brain, and consists of sensors, associative and responsive components.

The main advantage of using quantitative characteristics for the purpose of micro-object classification is the lack of a subjective human factor. Therefore, an urgent problem is evaluation of the quantitative characteristics of micro-objects, their reduction and data classification using modern classifiers to improve diagnostic accuracy.

2 Literature review

Classification is an important part of the data analysis process, which can be performed by different algorithms divided into different groups. These groups are based

on machine learning techniques [6]. Modern approaches to detection and classification of cell nuclei in cytological images are considered in [7]. In this paper, the authors compare classification results obtained using manual markups and deep learning methods. In [8], the authors provide a comparative analysis of the results of cytological image classification using the k-nearest neighbor method and the support vector method. In the experiments, the shape of the nuclei and the structure of the tissue were taken into account. The study of the k-nearest neighbor method is relevant in the field of data mining and machine learning [9]. Zhao in [10] developed a new algorithm based on the use of labeled samples. These methods were used mainly for fast searching [11], reducing the dimension, and improving the efficiency of algorithms. The support vector method is a set of learning methods used for classification and regression. They belong to the family of generalized linear classification [12].
In [13], a hybrid method of combining the support vector method and accelerator methods was developed. The authors show the effectiveness of the classification, and the SVM is used as the basic classifier for the data group classification. In [14-15], structures of convolutional neural networks were proposed for the classification of breast cancer histopathology images regardless of their degree of enlargement. The advantage of the developed systems on the basis of the proposed structures is the automation of the diagnostic process and the formation of a database for further research. Comparison of the quality of the breast cancer detection using magnetic resonance imaging and immunohistochemical studies is presented in [16]. A comparative analysis of approaches to biomedical image analysis is presented in [17,18].

The analysis of the above-mentioned publications has shown that scientists pay considerable attention to the problem of finding the ways of diagnosing precancerous and cancerous conditions of the breast on the basis of artificial intelligence systems. However, the complexity of the study and the large number of classifiers require additional research and comparative analysis, and a considerable amount of data needs to be reduced in size.

3 Problem statement

The purpose of this work is to analyze the existing means of artificial intelligence and their applications for the classification of cytological images based on the quantitative characteristics of micro-objects. To achieve this goal, the following tasks must be accomplished:

1. To reduce data using the principal component method.
2. To develop the structure of classification system of cytological images.
3. To conduct computer experiments in order to carry out a comparative analysis of the classifiers.

Formally, the formulation of the problem is as follows. Assume a set of features (1):

$$X = \{x_{ij}\}\ ,\ i = \overline{1,m}\ ,\ j = \overline{1,n}\ , \tag{1}$$

where n – a number of features, m – a number of their implementations.
In addition, a set of classes P is given.
To carry out the classification procedure we use a set of classifiers (2):

$$C = \{C_1, C_2, ..., C_t\}, \quad (2)$$

where t – a number of classifiers.

After applying the PCA, we obtain a set of components $Y = \{y_i\}$, where $Vi = \overline{1, m}$. Each component is a linear combination of features (3):

$$y_j = w_{1j}x_1 + w_{2j}x_2 + ... + w_{rj}x_r, \quad (3)$$

when $r < m$.

Thus, we get a set of components with their contributions $\{y_k, I_k\}$, $k = \overline{1, s}$, where s – a number of principal components.

The classification accuracy ε_0 is specified. Then, it is necessary to find s value in such a way:

$$s = \left|Y^s\right| = \min_{Y^s \in Y, \varepsilon \le \varepsilon_0} \left|Y\right|. \quad (4)$$

4 Principal component analysis

Input feature matrix X is given:

$$X = \begin{bmatrix} x_{11} & x_{12} & ... & x_{1j} & ... & x_{1n} \\ x_{21} & x_{22} & ... & x_{2j} & ... & x_{2n} \\ \vdots & \vdots & ... & \vdots & ... & \vdots \\ x_{i1} & x_{i2} & ... & x_{ij} & ... & x_{in} \\ \vdots & \vdots & ... & \vdots & ... & \vdots \\ x_{m1} & x_{m2} & ... & x_{mj} & ... & x_{mn} \end{bmatrix}. \quad (5)$$

In columns there are features, they are indexed by j ($j = \overline{1,n}$), and the lines are their implementations. They are indexed by i ($i = \overline{1,m}$).

The PCA implementation can be presented by a number of steps.

1. Centering and rationing of the output data is performed according to the formula (6):

$$\frac{x_{ij} - \overline{x}_j}{\sigma_j}, \tag{6}$$

where j – the number of the original variable, i – the implementation number of the j-*th* variable, and $\overline{x}_j$ and σ_j – the arithmetic mean and root mean square deviation of the x_j feature.

2. Calculating the covariance (S) or correlation (R) matrix.

$$S = \begin{bmatrix} \sigma^2_{x_1} & \text{cov}_{x_1,x_2} & \cdots & \text{cov}_{x_1,x_n} \\ \text{cov}_{x_2,x_1} & \sigma^2_{x_2} & \cdots & \text{cov}_{x_2,x_n} \\ \vdots & \vdots & \cdots & \vdots \\ \text{cov}_{x_n,x_1} & \text{cov}_{x_n,x_2} & \cdots & \sigma^2_{x_n} \end{bmatrix}, \tag{7}$$

$$\text{cov}_{x_k,x_j} = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ik} - \overline{x}_k\right)\left(x_{ij} - \overline{x}_j\right), \tag{8}$$

$$r_{x_k,x_j} = \frac{\text{cov}_{x_k,x_j}}{\sigma_{x_k}\sigma_{x_j}} \tag{9}$$

3. Finding the eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$ of the matrix S (or R) using characteristic equation (10):

$$\det\left(S - \lambda E\right) = 0 \text{ or } \det\left(R - \lambda E\right) = 0, \tag{10}$$

where E – a unitary matrix (a square matrix with ones on the diagonal and zeros elsewhere).

4. Finding the eigenvector for each eigenvalue λ_j. The eigenvector is the solution to the system of equations (11):

$$\left(S-\lambda E\right)\cdot\vec{w}=0\text{, or }\left(R-\lambda E\right)\cdot\vec{w}=0\text{,} \tag{11}$$

where $\vec{w}$ – eigenvector.

5. Finding linear combinations for principal components y_j

$$y_j = w_{1j}x_1 + w_{2j}x_2 + ... + w_{pj}x_p\text{.} \tag{12}$$

6. Analysis of the contribution of each of the principal components and their ranking in ascending order.

The contribution of each component is evaluated by the formula (13):

$$I_j = \frac{\lambda_j}{\lambda_1 + \lambda_2 + ... + \lambda_p}\text{,} \tag{13}$$

where j – a number of a component.

5 Structure of the cytological image classification module

The lack of clear contours of cell nuclei leads to difficulties in cytological image processing. Transmission of the digital image from the camera to the microscope and via communication channels to the computer causes pulse noises, which can often be classified by the software system as a part of the investigated object. Examples of cytological images of precancerous and cancerous conditions of the breast are taken from the BPCI2100 database [19](Fig. 1).

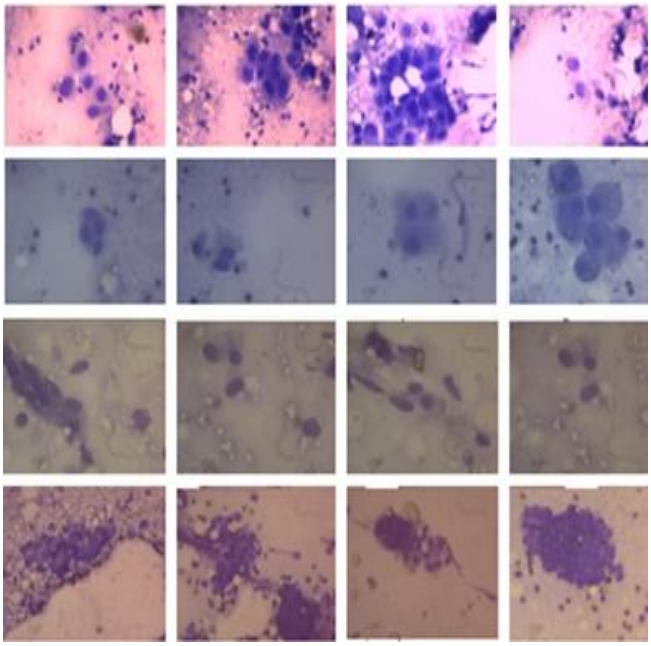


Fig. 1. Cytological images

During the study, the following quantitative characteristics of cell nuclei were determined: area, perimeter, length, width, circumference, coordinate Xc, coordinate Yc, length of the major axis, length of the minor axis, angle of inclination of the major axis to the OX axis, perimeter of a bounding box, coordinate Bx, coordinate By, bounding box width, bounding box length, bounding box area, aspect ratio.

The structure of cytological image classification module is shown in Figure 2.

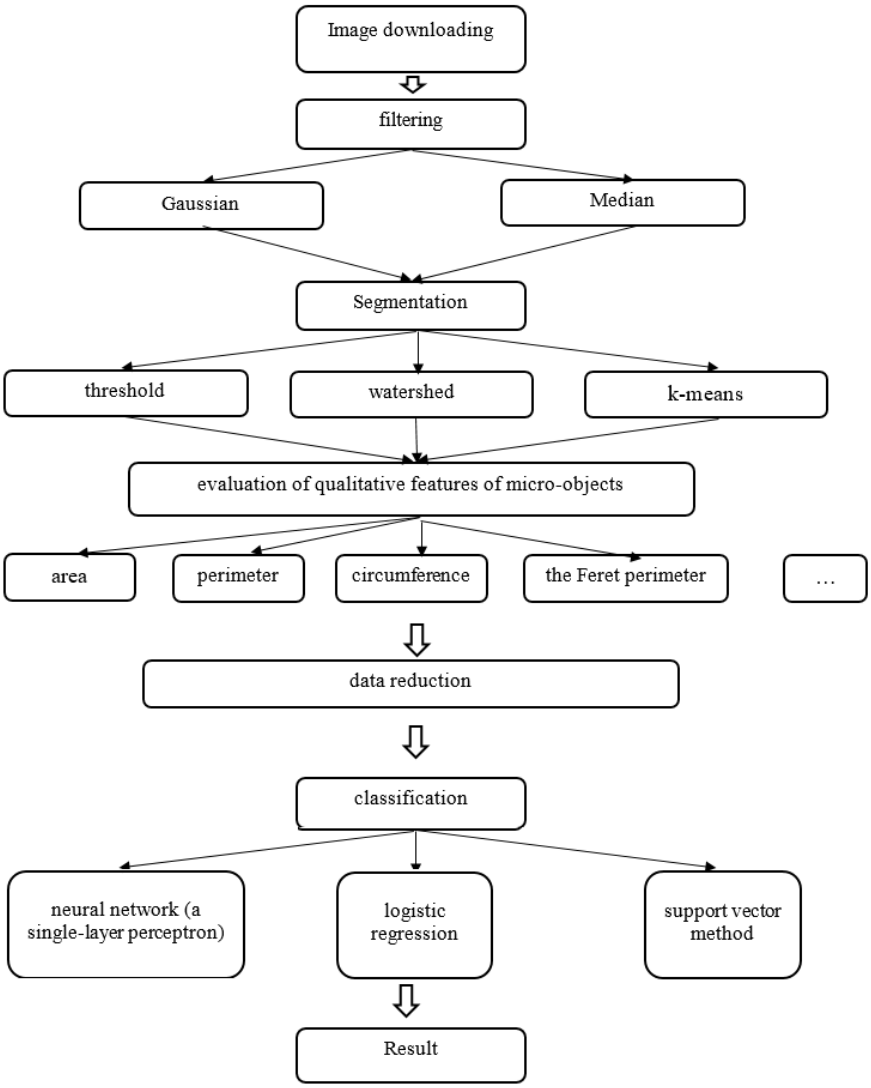


Fig. 2. Structure of cytological image classification module

The classification process consists of the following steps:

1. Filtering the input image. This stage makes it possible to reduce noise. Gaussian and Median filtering is used to reduce Gaussian and pulse noise, respectively. Filtering can significantly improve image quality that will have positive impact on the next stages.

2. Segmentation. A segmentation stage is required to select particular areas in the image (cell, background nuclei). For cytological images, the best results were shown by watershed algorithms, k-means, and threshold segmentation. Based on these algorithms, an algorithm for segmentation of cytological and histological images was

developed [20]. Quantitative evaluation of image segmentation quality was performed on the basis of Gromov-Frechet and Gromov-Hausdorff metrics.[21]

Cell nuclei are selected using the image contour selection algorithm developed by the authors[22,23].

After converting the image into the type of "white background – black objects", the quantitative characteristics of the particular micro-objects are evaluated.

3. Data reduction. The principal component method is used to reduce the data size.

4. Classification. Data in the form of an array of numbers is fed to the classification module. The classification module implements the support vector method, logistic regression, a single-layer perceptron and the k-nearest neighbor method. Classification results in an associative array that includes data on the parameters of the nuclei and their corresponding classes.

To analyze the classification results, ROC-curves were constructed and AUC coefficients were calculated.

6 Structure of the cytological image classification module

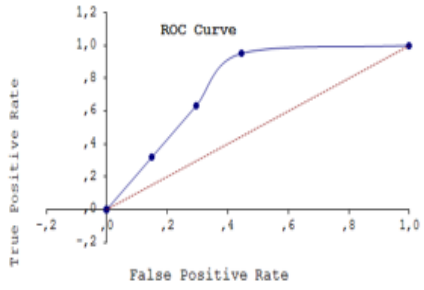
The software module for testing the cytological image classification techniques is written in Java programming language and deeplearning4j library.

An example of the data reduction of the analysis of the cytological image cell nuclei is given in Figure 3.

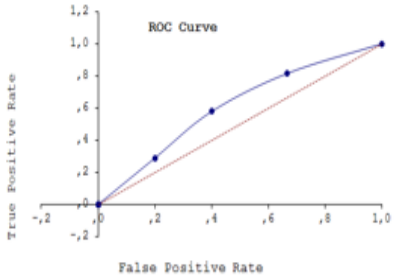
Variable	**Factor 1**	Factor 2	Factor 3
area	0,930062	0,115608	0,202682
perimeter	0,965469	0,166686	-0,055015
length	0,753929	-0,259835	-0,288082
width	0,778245	0,393526	0,142044
circle circle circumference	0,974332	0,102113	0,116891
coordinate Xc	-0,562509	0,543321	0,114089
coordinate Yc	0,105962	0,824972	-0,210445
length of the major axis	0,870384	0,233751	-0,327477
length of the minor axis	0,862471	-0,028791	0,472105
angle of inclination of the major axis to the OX axis	-0,337040	0,440998	0,663066
perimeter of a bounding box	0,977713	0,123953	0,084530
coordinate Bx	-0,569342	0,540003	0,112482
coordinate By	0,095465	0,829533	-0,209305
bounding box width	0,923290	0,165663	0,124006
bounding box length	0,718818	-0,319073	-0,088216
rectangle area	0,902130	-0,135494	0,072055
aspect ratio	0,076785	0,203315	-0,905569
Expl.Var	9,341294	2,683717	1,900034
Prp.Totl	0,549488	0,157866	0,111767

Fig. 3. Data after reduction (3 main components)

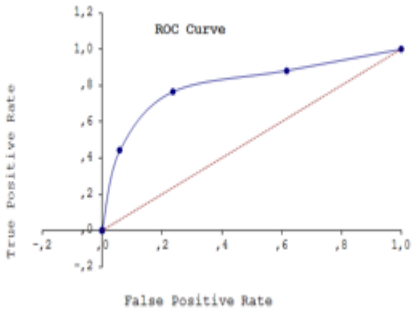
The results of the classification of the cytological images by the support vector method are shown in Figure 4.



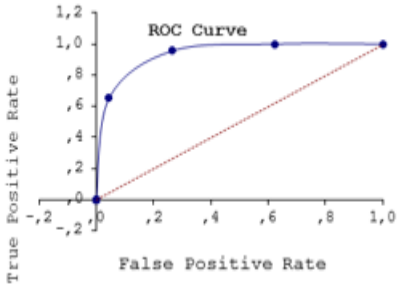
2 components (AUC = 0.75)



3 components (AUC = 0.6



5 components (AUC = 0.79)



8 components (AUC = 0.91)

Fig. 4. The results of the classification of the cytological images by the support vector method

Classification quality score using the support vector method on the basis of the AUC coefficient is shown in Figure 5.

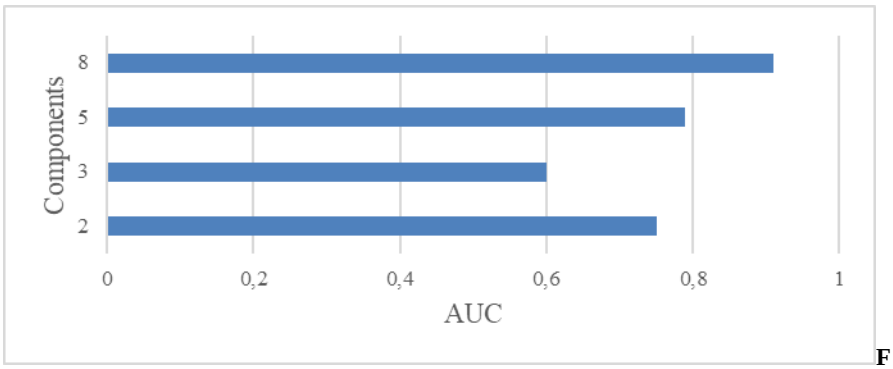
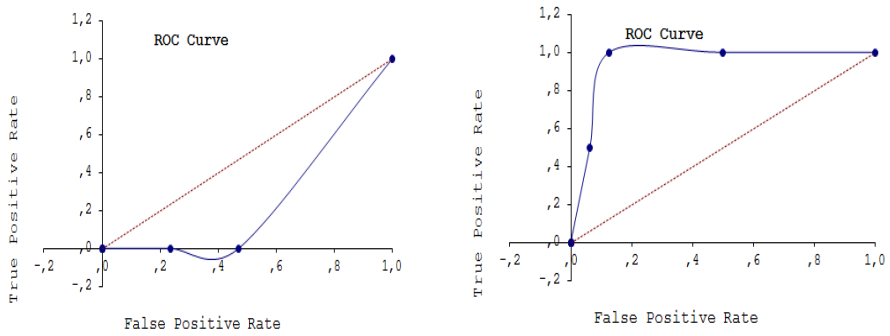


Fig. 5. Classification quality score using the support vector method

The analysis of the results in Figure 5 shows that the highest classification accuracy can be achieved using 8 components. Classification results with the use of a single-layer perceptron are shown in Figure 6.



2 components (AUC=0.27) 3 components (AUC=0.93)

Fig. 6. Classification results with the use of a single-layer perceptron

The AUC coefficient shows that the best classification quality is obtained with the use of 3 components.

The results of the classification of the cytological images based on multiclass logistic regression are shown in Figure 7.

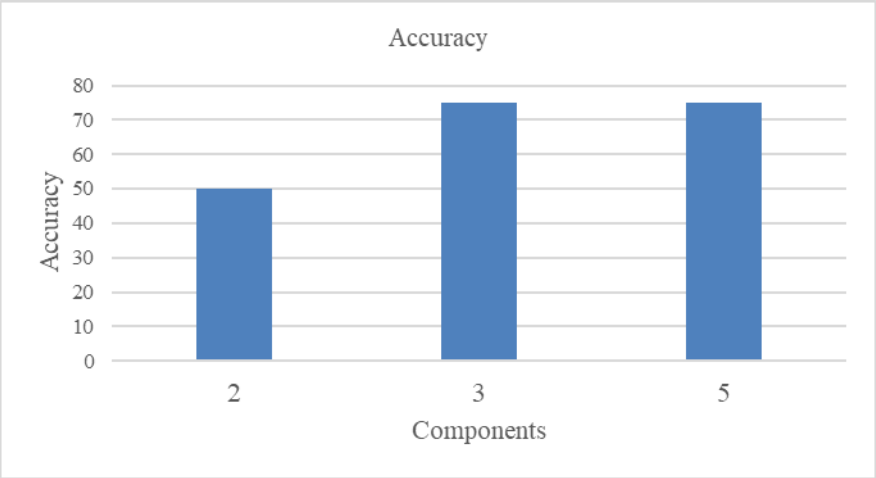


Fig. 7. The results of the classification of the cytological images based on multiclass logistic regression

Therefore, the best result is achieved with the use of 3 and 5 components and is 75%.

Conclusions

1. Using the basic algorithms of low, medium and high levels of computer vision, a classification structure of cytological images is developed. The developed structure includes the use of the principal component method to reduce the data size.
2. Applying the principal component method, the input indicators were reduced which showed that mainly three components are informative.
3. Computer experiments have shown that the best result of the classification of the quantitative characteristics of the cytological image nuclei is obtained using the 3 major components. The AUC is about 93%.

References

1. Tsmots I., Teslyuk V., Teslyuk T., Ihnatyev I. Basic Components of Neuronetworks with Parallel Vertical Group Data Real-Time Processing. In: Shakhovska N., Stepashko V. (eds) Advances in Intelligent Systems and Computing II. CSIT 2017. Advances in Intelligent Systems and Computing, Springer, Cham, vol. 689, 558 – 576 (2017).
2. Berezsky O., Pitsun O., Batryn N., Berezska K., Dubchak L. Modern automated microscopy systems in oncology. Proceedings of the 1st International Workshop on Informatics & Data-Driven Medicine, Lviv, Ukraine, 28-30 november 2018 – p. 16 (2018)

3. Oliinyk A.O., Subbotin S. O., Oliinyk O.O. Intelektualnyi analiz danykh : navchalnyi posibnyk . – Zaporizhzhia : ZNTU. – 278 (2012)
4. Suthaharan S. Support Vector Machine. Machine Learning Models and Algorithms for Big Data Classificationю – pp. 207-235 (2016)
5. Agresti A. Categorical Data Analysis. Wiley-Interscience, New Jork (2002)
6. Zhang Z. Introduction to machine learning: k-nearest neighbors. Ann Transl Med. 2016 Jun; 4(11), 218 (2016)
7. Trifonov R., Gotseva D., Angelov V. Binary classification algorithms. International Journal of Development Research. 7 (11), 16873-16879 (2017)
8. Hady Ahmady Phoulady, Peter R. Mouton. A New Cervical Cytology Dataset for Nucleus Detection and Image Classification (Cervix93) and Methods for Cervical Nucleus Detection. https://arxiv.org/abs/1811.09651, last accessed 2019/10/28
9. Loukas C., Kostopoulos S., Tanoglidi A., Glotsos D. Breast Cancer Characterization Based on Image Classification of Tissue Sections Visualized under Low Magnification Computational and Mathematical Methods in Medicine. Volume 2013, p. 7 (2013) http://dx.doi.org/10.1155/2013/829461
10. Zhu X., Huang Z., Cheng H., Shen H. Sparse hashing for fast multimedia search ACM Trans. Inform. Syst. 31 (2), 9:1-9:24 (2013)
11. Zhao D. Zou W., Sun G.. A fast image classification algorithm using support vector machine. 2nd International Conference on Computer Technology and Development . Egypt, (2010) DOI: 10.1109/ICCTD.2010.5645823
12. Zhang S. Efficient kNN classification algorithm for big data. Neurocomputing. 195 143–148 (2016).
13. Dhivyapriya P., Sivakumar P. Classification of Cancer Dataset in Data Mining Algorithms Using R Tool. International Journal of Computer Science Trends and Technology (IJCST), 5 (1) (2017)
14. Sharma A., Dey S., A boosted SVM based ensemble classifier for sentiment analysis of online reviews, Appl. Comput. Rev. 13, 43–52 (2013)
15. Bayramoglu N, Kannala J, Heikkila J. Deep Learning for Magnification Independent Breast Cancer Histopathology Image Classification. 23rd International Conference on Pattern Recognition (ICPR). (2016) DOI: 10.1109/ICPR.2016.7900002
16. Berezsky O., Pitsun O., Verbobyy S., Datsko T., Bodnar A. Computer diagnostic tools based on biomedical image analysis. 14th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM), 21-25 Feb. 2017, Lviv, Ukraine, 388 – 391 (2017)
17. Bae M.S. Quantitative MRI morphology of invasive breast cancer: correlation with immunohistochemical biomarkers and subtypes / Min Sun Bae, Mirinae Seo Mirinae Se// Acta Radiol. 2015 Mar;56(3):269-75. doi: 10.1177/0284185114524197
18. Berezsky, O., Pitsun, O., Verbovyy, S., Datsko, T., Bodnar, A. Computer diagnostic tools based on biomedical image analysis. 14th International Conference The Experience of Designing and Application of CAD Systems in Microelectronics, CADSM 2017, - pp. 388 – 391 (2017)
19. Berezsky O., Melnyk H., Verbovyi S., Pitsun O., Nykoliuk V., Datsko T. Certificate of copyright registration № 75359. Database of digital cytological and histological images of the precancerous and cancerous conditions of the breast "BPCI2100". Date of registration 12.14.2017.
20. Berezsky O., Batko Yu., Melnyk G., Verbovyy S. Segmentation of Cytological and Histological Images of Breast Cancer Cells. IEEE 8th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), Warsaw, Poland (2015) doi: 10.1109/IDAACS.2015.7340745
21. Berezsky, O., Zarichnyi M. Gromov-Fréchet distance between curves. Matematychni Studii, 50 (1), 88-92 (2018).

22. Berezsky O., Bat'ko Yu. Algorithm of determination of image contours of biological nature. Proceedings of the International conference «Modern problems of radio engineering, telecommunications and computer science» TCSET2006. Lviv-Slavske. – Lviv, pp. 642-644 (2006)
23. Berezsky O., Verbovyy S., Pitsun O. Hybrid intelligent information technology for biomedical image processing. IEEE 13th International Scientific and Technical Conference on Computer Sciences and Information Technologies, CSIT 2018 – Proceedings, Lviv, Ukraine (2018).