Improving the quality of cytological cell nuclei classification using ensembles

Oleh Pitsun[*,†] and Myroslav Shymchuk[†]

West Ukrainian National University, 11 Lvivska st., Ternopil, 46001, Ukraine

Abstract

Biomedical image classification is an important point of any automated diagnostic system. The disadvantage of existing studies is the use of a limited number of machine learning algorithms for classification, data segmentation. The impact of image quality on classification and especially segmentation results is particularly noticeable when processing biomedical images. Since the nuclei of cells in biomedical images are characterized by the complexity of processing, the use of one or two algorithms for all types of images is insufficient. In this work, modern machine learning algorithms with and without a teacher are used to classify the quantitative characteristics of cell nuclei. In this work, the use of an ensemble approach with a combination of several algorithms and the use of the soft, hard voting principle is proposed. The proposed solution allowed to obtain an accuracy of 99.22% for the training sample and 83.12% for the test sample.

Keywords

CNN, images, parallel processing

1. Introduction

Methods for classifying biomedical images are key to modern medical diagnostics, since classification can increase the automation of processes and increase the accuracy of detecting pathological changes in cells. However, the efficiency of classification of individual models may decrease due to the influence of noise, artifacts, or significant variability of the data. To solve these problems, ensemble machine learning methods, such as Random Forest, Gradient Boosting, etc., are increasingly used.

The relevance of suction is due to the ability to combine the results of many weak models into a mixed hybrid predictor. This approach provides increased accuracy and reliability of classification. Also, the use of several algorithms provides a reduction in errors and resistance to changes in data. The use of such methods in the classification of cytological, histological, immunohistochemical images is a promising direction for improving automatic diagnostics and decision support in biomedical research and clinical practice.

Calculation of quantitative characteristics of cell nuclei on immunohistochemical, histological and cytological images is relevant, as it allows for an objective assessment of morphological changes accompanying pathological processes. The main parameters that describe cell nuclei are: area, perimeter, circumference, length of the major and minor axes. Automated quantitative analysis increases the objectivity of the assessment and contributes to the development of computer-aided decision-making systems in medical practice. Another advantage of using such an approach to diagnosis is the absence of the need to process large-sized images with a large noise component and erroneous data. The use of quantitative characteristics for classification allows you to speed up the classification process and avoid the need for large processing resources. Hard voting and soft voting algorithms play an important role in increasing the accuracy and reliability

[*] *CIAW-2025: Computational Intelligence Application Workshop, September 26-27, 2025, Lviv, Ukraine*

[1*] Corresponding author.

[†] These authors contributed equally.

o.pitsun@wunu.edu.ua (O. Pitsun); Miroslavshimchuk@gmail.com (M. Shymchuk)

0000-0003-0280-8786 (O. Pitsun); 0009-0009-4270-8131 (M. Shymchuk)

of classification in machine learning tasks. Their use allows for more efficient combination of the results of a specific set of basic models, which in turn provides more accurate forecasting.

The object of the study is the quantitative characteristics of biomedical image kernels.

The subject of the study is an ensemble method for classifying quantitative characteristics, which uses hard voting and soft voting.

The purpose of this work is to develop an ensemble method for classifying quantitative characteristics of biomedical image kernels. A feature of the proposed work is the use of the PCA algorithm to increase the accuracy of classification in combination with the ensemble method.

2. Problem statements

The following tasks were performed in this work:

1. an analysis of machine learning algorithms for classifying quantitative characteristics of objects was performed;
2. an analysis of approaches to calculating quantitative characteristics of cell nuclei in biomedical images was performed;
3. a comparative analysis of machine learning algorithms using ensemble methods to obtain the best result was performed.

3. Literature review

In [1], the architecture of the ensemble framework for data classification is presented to improve the quality of classification. The scope of application of this framework is medicine. Many data in medicine are unbalanced, so in [2] the Undersampling Balanced Ensemble (USBE) algorithm is proposed. As a result, the authors achieved better data classification performance using two different breast cancer datasets. Quantitative characteristics are the basic element for classification tasks, however, in [6] an approach is proposed that allows the use of convolutional neural networks as one of the classification methods. The proposed approach improves accuracy compared to SVM. In [4] the authors developed an ensemble method that is designed for classifying medical datasets.

In [6] the authors present a review and analysis of modern ensemble methods for forecasting. The work also clearly highlights the problems and trends in this field. In [9], a method is proposed that aims to minimize the subset of features to achieve a satisfactory diagnosis of a wide range of diseases with the highest accuracy, sensitivity and specificity. The ensemble method is also used in the analysis of medical datasets.

In [7], the authors tested the effectiveness of the proposed ensemble learning method on nine unbalanced medical datasets. The experimental results showed that this paradigm outperforms other modern classification models. In [0], the authors classify the quantitative characteristics of cell nuclei and use different approaches to classify medical data. The authors also propose to use existing approaches for further research in other fields than medicine. In the study [9], the raw data set was first pre-processed, then a machine learning method was applied, including artificial neural network, decision tree, support vector machine, naive Bayes, and nearest neighbor (KNN) with one ensemble method (which collects 30 KNN algorithms as weak learners). The prediction result was obtained using the majority vote method based on the generator output data.

The work [10] emphasizes the importance of obtaining and comparing the performance of different types of ensemble machine learning models during electronic medical record screening. Papers [11-13] present an analysis of modern approaches to the analysis of cell nucleus characteristics, which allows us to highlight their features. In paper [14-18], modern techniques for using ensemble methods for classifying images of cell nuclei, in particular cytological and histological ones, were analyzed. The analysis performed allows us to determine the main algorithms for ensembles. Application of mlops practices for biomedical image classification shown in [19].

The above analysis demonstrates the relevance of the application of ensemble methods in medicine, in particular in the tasks of classifying data in the form of quantitative characteristics of cell nuclei.

4. Materials and methods

Calculation of quantitative characteristics of cell nuclei

To obtain information about the quantitative characteristics of cell nuclei, image processing was performed using preprocessing and segmentation. The main characteristics of cells are:

1. area;
2. perimeter;
3. length of the main and lateral axes;

Figure 1 shows an example of calculating the quantitative characteristics of cell nuclei.

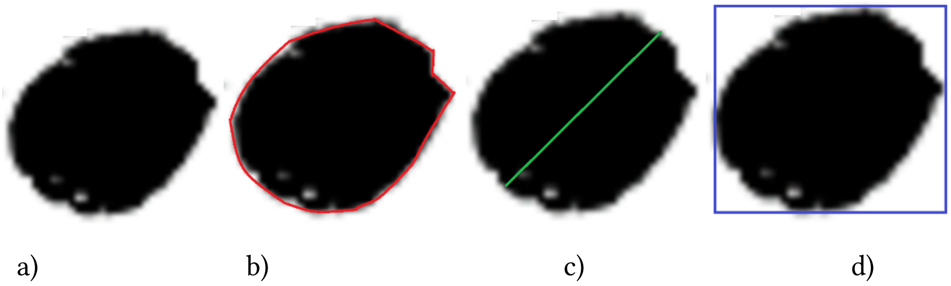
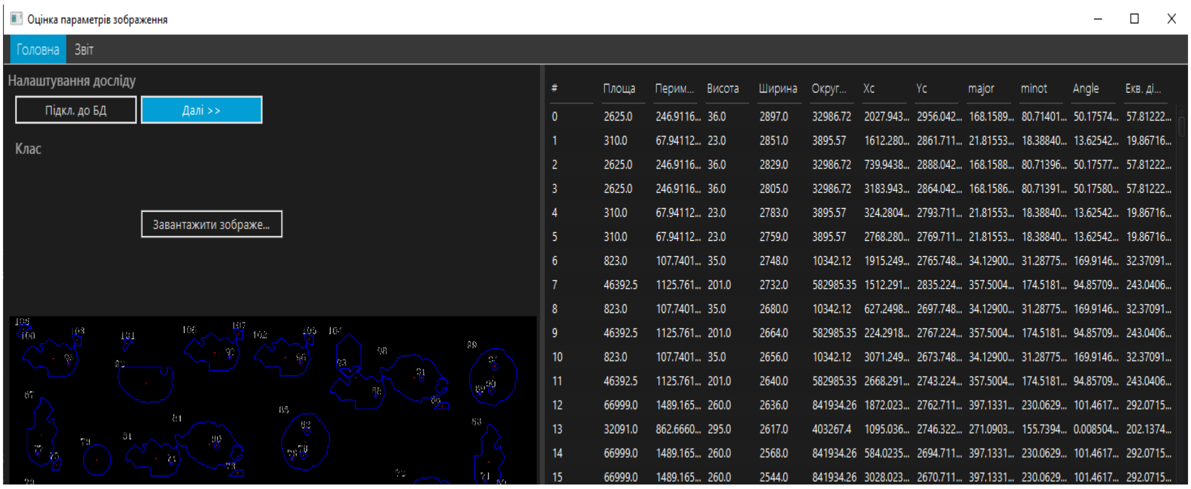


Figure 1: Visualization of the process of calculating quantitative characteristics (a – original image, b – area, c – main axis, d – area of bounding rectangle).

The interface of the software module for calculating quantitative characteristics is shown in Figure 2.



#	Площа	Перим...	Висота	Ширина	Округ...	Xc	Yc	major	minot	Angle	Екв. ді...
0	2625.0	246.9116...	36.0	2897.0	32986.72	2027.943...	2956.042...	168.1589...	80.71401...	50.17574...	57.81222...
1	310.0	67.94112...	23.0	2851.0	3895.57	1612.280...	2861.711...	21.81553...	18.38840...	13.62542...	19.86716...
2	2625.0	246.9116...	36.0	2829.0	32986.72	739.9438...	2888.042...	168.1588...	80.71396...	50.17577...	57.81222...
3	2625.0	246.9116...	36.0	2805.0	32986.72	3183.943...	2864.042...	168.1586...	80.71391...	50.17580...	57.81222...
4	310.0	67.94112...	23.0	2783.0	3895.57	324.2804...	2793.711...	21.81553...	18.38840...	13.62542...	19.86716...
5	310.0	67.94112...	23.0	2759.0	3895.57	2768.280...	2769.711...	21.81553...	18.38840...	13.62542...	19.86716...
6	823.0	107.7401...	35.0	2748.0	10342.12	1915.249...	2765.748...	34.12900...	31.28775...	169.9146...	32.37091...
7	46392.5	1125.761...	201.0	2732.0	582985.35	1512.291...	2835.224...	357.5004...	174.5181...	94.85709...	243.0406...
8	823.0	107.7401...	35.0	2680.0	10342.12	627.2498...	2697.748...	34.12900...	31.28775...	169.9146...	32.37091...
9	46392.5	1125.761...	201.0	2664.0	582985.35	224.2918...	2767.224...	357.5004...	174.5181...	94.85709...	243.0406...
10	823.0	107.7401...	35.0	2656.0	10342.12	3071.249...	2673.748...	34.12900...	31.28775...	169.9146...	32.37091...
11	46392.5	1125.761...	201.0	2640.0	582985.35	2668.291...	2743.224...	357.5004...	174.5181...	94.85709...	243.0406...
12	66999.0	1489.165...	260.0	2636.0	841934.26	1872.023...	2762.711...	397.1331...	230.0629...	101.4617...	292.0715...
13	32091.0	862.6660...	295.0	2617.0	403267.4	1095.036...	2746.322...	271.0903...	155.7394...	0.008504...	202.1374...
14	66999.0	1489.165...	260.0	2568.0	841934.26	584.0235...	2694.711...	397.1331...	230.0629...	101.4617...	292.0715...
15	66999.0	1489.165...	260.0	2544.0	841934.26	3028.023...	2670.711...	397.1331...	230.0629...	101.4617...	292.0715...

Figure 2: Graphical interface for calculating quantitative characteristics of cell nuclei.

After the calculations are performed, the results are stored in a csv file for further processing by the classifier.

Classification Algorithms

Ensemble learning helps improve the performance of a machine learning model by combining multiple models. This approach allows for better prediction performance compared to a single model. The main causes of training errors are noise, bias, and variance. Ensemble helps minimize these factors.

Bagging is the application of the Bootstrap procedure to a high-variance machine learning algorithm, typically a decision tree. Boosting is a sequential process where each subsequent model tries to correct the errors of the previous model. Subsequent models depend on the previous model.

Bagging and Boosting reduce the variance of a single estimate because they combine multiple estimates from different models.

Logistic Regression

The essence of regression analysis is to analytically or experimentally determine the coefficients of the features of objects in such a way as to minimize the total error between the values of the model function and the experimental data for the entire input sample.

Logistic regression is a type of multiple regression used to study the relationship between a binary or categorical outcome and several influencing factors. The paper [1] demonstrates the basic principle, the choice and role of the independent variable, the conditions of application, the model estimation and diagnostics of multiple logistic regression. The goal of regression is to determine whether an object belongs to one of two classes, where a set of object features represents the input variable, and the output variable (analysis result) is binary and takes the values 0 or 1. The advantages of logistic regression include:

1. Ease of implementation;
2. High efficiency of working with large data sets;
3. Relatively high quality of working with unbalanced data.

The objective function of linear regression is defined as the minimization of the mean square error between the predicted and actual values of the observed parameter [2]:

$$F=\frac{1}{n}\sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2 \rightarrow min \tag{1}$$

where $\hat{y}_i$ – predicted (analytical) value, linear concerning the desired coefficients;

y_i – the actual value;

n – data sample length.

General linear regression model:

$$\hat{y}=w_0+w_1x_1+w_2+\ldots+w_mx_m=\sum_{i=0}^{m} w_ix_i=W^T X \epsilon R^I, \tag{2}$$

where $W^T=\left(w_0, w_1, \ldots, w_m\right)$ - weights of object features;
$w_i \epsilon R^I$; X=(1,x_1,$x_2, \ldots, x_m$) – object predictors;
m – number of features of the object; T – transposition symbol.

Random Forest

Random forest is an ensemble machine learning method used to solve classification, regression, and other types of prediction problems. Its working principle is to create a set of decision trees during model training, after which a class is selected for classification based on the majority of votes. For regression, the average value of the predictions of all trees is calculated. The algorithm is given in [3]

Algorithm for building a committee:

1. Generating a random subsample with repetition of size *n* from the training sample;
2. Randomly select *m* predictors (features) from *M;*

3. The decision tree is constructed by selecting the features based on which the partition is performed, not from all *M* features, but only from *m* randomly selected ones;
4. Dividing the trait *X* into two classes $X_i \geq S_i$ and $X_i < S_i$;
5. Measure homogeneity in two new classes using the Gini criterion;
6. Take the following value of the "split point" S_i feature *X*, for which maximum class homogeneity is achieved;
7. The tree is built until the subsample is fully used without applying the pruning procedure;
8. Returning to step 1, generate a new sample and repeat steps 2–4 to build the next tree.

Classification of objects is carried out by voting: each tree in the ensemble assigns the object to one of the classes, and the class that received the most significant number of votes from the trees is chosen.

Gradient Boosting

Gradient Boosting is a machine learning technique used to solve classification and regression problems. The basic idea is that a collection of weak models can produce a more accurate predictor when working together. Gradient boosting combines several weak models to form a single strong model. Although algorithms with fast learning can be challenging to optimize, their accuracy is significantly improved by sequential coupling. In contrast, models with a slower learning curve adapt better to statistical patterns in the data. Weak learners are added in such a way that each subsequent learner takes into account the residuals obtained in the previous stage during the model development process. The final model combines the results of all stages, forming a strong learner. The residuals are calculated using a loss function.

Hard voting

Hard voting is a simple algorithm that can be used to combine predictions from multiple classifiers. The algorithm works by first predicting each classifier. The ensemble prediction is then simply the majority vote of the individual classifiers.

The simplicity of hard voting makes it a popular choice for machine learning practitioners. Hard voting is also very efficient and can often outperform other ensemble learning algorithms. Finally, hard voting can be used with any classifier, making it a versatile tool. Hard voting can also be less robust to noise in the data. This is because hard voting is based on the majority votes of the individual classifiers. If the data is noisy, it is more likely that the majority of votes will be wrong.

Predict the class label $\hat{y}$ by the majority vote of each classifier C_j

$$\hat{y} = mode\{C_1(x), C_2(x), \ldots, C_m(x)\} \tag{3}$$

For example, if two classes provided zero and one class provided 1:

$$\hat{y} = mode\{0, 0, 1\} = 0 \tag{4}$$

In addition to simple majority voting as described above, a weighted majority vote can be calculated by relating the weight:

$$\hat{y} = arg\,max \sum_{j=1}^{m} w_j X_A(C_j(x) = i) \tag{5}$$

where X_A is the characteristic function $[C_j(x) = i \epsilon A,]$and A is the set of unique class labels

Soft Voting

Soft voting is an algorithm that can be used to combine the predictions of multiple classifiers based on their probabilities. The algorithm works by first assigning a probability to each class. The ensemble prediction is then simply the class with the highest overall likelihood.

Soft voting has several advantages over hard voting. First, it is more accurate than hard voting. Second, it is more robust to noise in the data. Third, it can be used with any classifier

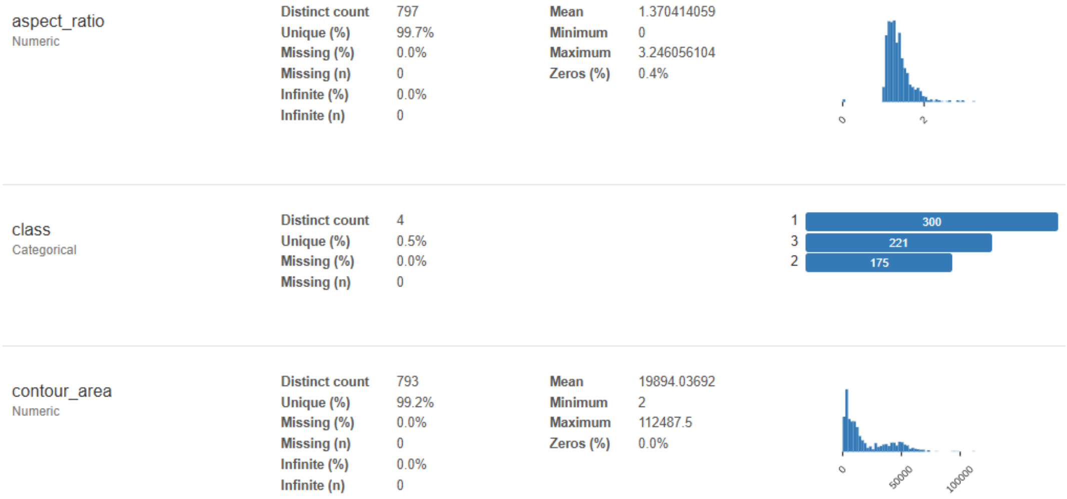
$$\hat{y} = arg\,max \sum_{j=1}^{m} w_j p_{ij} \tag{6}$$

where w_j is the weight that can be assigned to the *jth* classifier.

Using uniform weights, we calculate the average probabilities

$$\hat{y} = arg\,max\left[p\left(i_0 | X\right), p\left(i_0 | X\right)\right] \tag{7}$$

A statistical description of the parameters of cytological examination of quantitative characteristics of cell nuclei is given in Figure 3.



aspect_ratio
Numeric

Distinct count	797	Mean	1.370414059
Unique (%)	99.7%	Minimum	0
Missing (%)	0.0%	Maximum	3.246056104
Missing (n)	0	Zeros (%)	0.4%
Infinite (%)	0.0%		
Infinite (n)	0		

class
Categorical

Distinct count	4
Unique (%)	0.5%
Missing (%)	0.0%
Missing (n)	0

1	300
3	221
2	175

contour_area
Numeric

Distinct count	793	Mean	19894.03692
Unique (%)	99.2%	Minimum	2
Missing (%)	0.0%	Maximum	112487.5
Missing (n)	0	Zeros (%)	0.0%
Infinite (%)	0.0%		
Infinite (n)	0		

Figure 3: Statistical description of the parameters of cytological study of quantitative characteristics of cell nuclei.

The correlation matrix of the input data is shown in Figure 4.

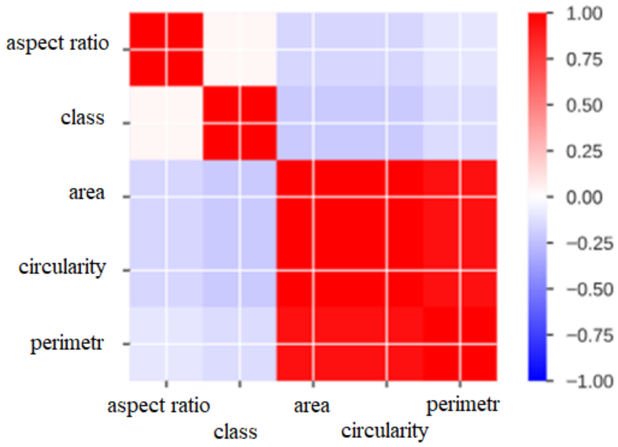


Figure 4: Input data correlation matrix.

Analyzing the above parameters, we can conclude that there is a need for data reduction to optimize classification performance and improve classification quality.

A quantitative description of the parameters is given in Table 1.

Table 1
Quantitative description of parameters

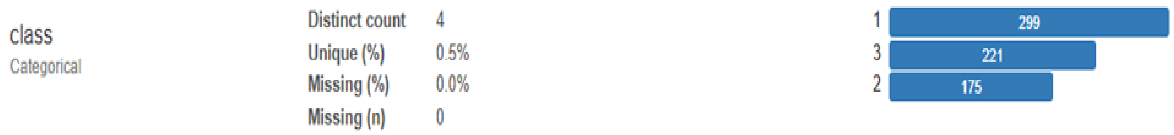
	Contour area	Contour perimeter	Contour circularity	Aspect ratio
Count	799	799	7	79
mean	19894	530	2	1.3
std	19711	350	2	0.3
min	2	5	2	0.0
25%	4326	261	5	1.1
50%	11166	416	1	1.3
75%	34461	756	4	1.4
max	112487	2115	1	3.2

The results of the classifiers without PCA are given in Table 2.

Table 2
The results of the classifiers without PCA

Model	Score train	Score test	Score diff
Stochastic Gradient Decent	27.39	28.75	1.36
Ridge Classifier	61.66	60.00	1.66
Perceptron	21.44	23.12	1.68
Lienar SVC	73.4	71.25	2.15
Neural Natwork	38.03	35.62	2.41
Ada Boost	3.56	56.88	2.58
Extra Trees	82.94	79.38	3.56
Logistic Regression	74.18	70.62	3.56
Naïve Bayes	75.59	71.88	3.71
LGBM	79.81	75.00	4.81
XGBC	96.24	82.5	13.74
kNN	86.38	72.5	13.88
Bagging	97.5	**83.12**	14.38
Voting soft	**99.06**	**83.12**	**15.94**
Voting hard	**99.06**	**82.5**	**16.56**
Decission Tree	99.84	79.38	20.46
SVM	99.84	36.25	63.59

The correlation between classes after PCA is shown in Figure 5.



class
Categorical

Distinct count	4
Unique (%)	0.5%
Missing (%)	0.0%
Missing (n)	0

Figure 5: Correlation between classes after PCA.

The above description allows us to evaluate the input data better and understand the range of values for each of the parameters after PCA processing.

The correlation matrix of the input data after PCA is shown in Figure 6.

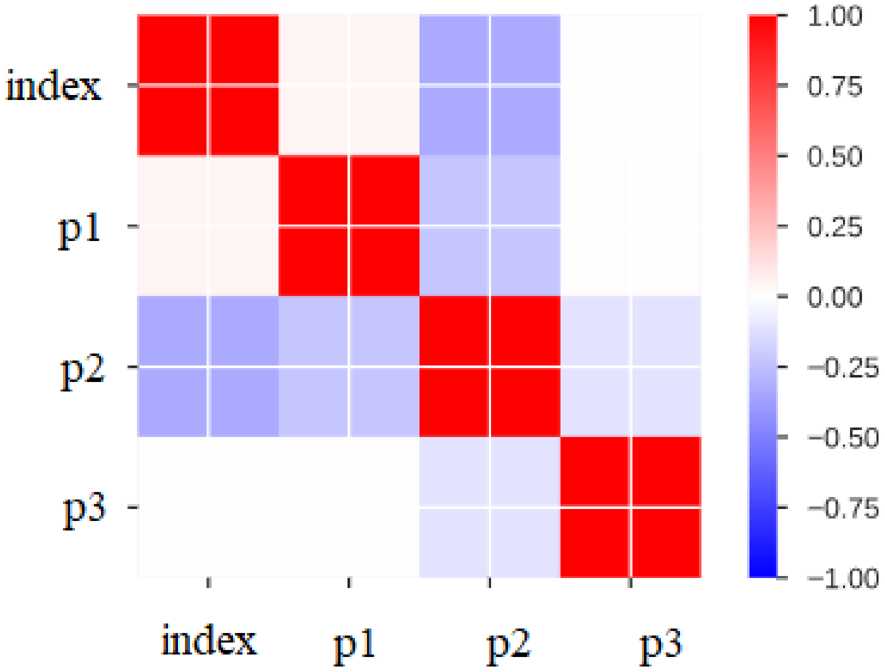


Figure 6:
Correlation matrix of input data after PCA.

The classification results after using data reduction are shown in Table 3.

Table 3
The results of the classifiers after using data reduction

Model	Score train	Score test
Stochastic Gradient Decent	58.31	56.25
Ridge Classifier	61.91	59.38
Perceptron	20.85	20.62
Lienar SVC	73.4	71.25
Neural Natwork	38.03	35.62
Ada Boost	47.35	51.75
Extra Trees	82.94	79.38
Logistic Regression	75.32	56.25
Naïve Bayes	78.12	75.00
LGBM	99.75	75.62
XGBC	95.3	78.15
kNN	86.38	72.5
Bagging	97.81	**83.12**
Voting soft	**97.02**	**78.75**
Voting hard	**99.86**	**82.25**
Decission Tree	99.84	80.0
SVM	99.84	36.25

Analyzing the indicators with and without PCA, it can be concluded that the Voting soft and complex algorithms demonstrated the best results.

Conclusions

As a result of the comparative analysis, it was found that the ensemble method based on the random_forest and bagging_classifier algorithms showed the best results compared to other classical data classification algorithms.

Based on the experimental approach, it was also found that the use of data reduction significantly increases the classification accuracy.

The classification accuracy using ensembles is 99.22% for the training sample and 83.12% for the test sample. The worst indicators were shown by approaches based on neural networks, the support vector method, and Adaboost.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] K. Firoz, V. Balusupati, S. Syed, I. Ashraf, L. Ramasamy. An Efficient, Ensemble-Based Classification Framework for Big Medical Data - Big Data - Vol. 10, No. 2 – 2022 https://doi.org/10.1089/big.2021.0132

[2] B. Krawczyk and G. Schaefer, "Ensemble fusion methods for medical data classification," 11th Symposium on Neural Network Applications in Electrical Engineering, Belgrade, Serbia, 2012, pp. 143-146, doi: 10.1109/NEUREL.2012.6419993

[3] L. Nanni, S. Brahnam, A. Loreggia, and L. Barcellona. 2023. "Heterogeneous Ensemble for Medical Data Classification" Analytics 2, no. 3: 676-693. https://doi.org/10.3390/analytics2030037

[4] L.R. Namamula, D. Chaytor, Effective ensemble learning approach for large-scale medical data analytics. Int J Syst Assur Eng Manag 15, 13–20 (2024). https://doi.org/10.1007/s13198-021-01552-7

[5] O. Sagi, and L. Rokach. "Ensemble learning: A survey." Wiley interdisciplinary reviews: data mining and knowledge discovery 8, no. 4 (2018): e1249. https://doi.org/10.1002/widm.1249

[6] Q. Al-Tashi, H. Rais and S. J. Abdulkadir, "Hybrid Swarm Intelligence Algorithms with Ensemble Machine Learning for Medical Diagnosis," 2018 4th International Conference on Computer and Information Sciences (ICCOINS), Kuala Lumpur, Malaysia, 2018, pp. 1-6, doi: 10.1109/ICCOINS.2018.8510615.

[7] N. Liu, X. Li, E. Qi, M. Xu, L. Li and B. Gao, "A Novel Ensemble Learning Paradigm for Medical Diagnosis With Imbalanced Data," in IEEE Access, vol. 8, pp. 171263-171280, 2020, doi: 10.1109/ACCESS.2020.3014362

[8] J.Sidey-Gibbons, C. Sidey-Gibbons, Machine learning in medicine: a practical introduction. BMC Med Res Methodol 19, 64 (2019). https://doi.org/10.1186/s12874-019-0681-4

[9] Z. Asghari Varzaneh, M. Shanbehzadeh & H. Kazemi-Arpanahi. Prediction of successful aging using ensemble machine learning algorithms. BMC Med Inform Decis Mak 22, 258 (2022). https://doi.org/10.1186/s12911-022-02001-6

[10] C. Stevens, A. Lyons, K. Dharmayat. Ensemble machine learning methods in screening electronic health records: A scoping review. DIGITAL HEALTH. 2023;9. doi:10.1177/20552076231173225

[11] M. Aghera, K. V. Singh, K. Vaishnani, U. Oza and B. Gohel, "Segmentation of Nuclei in H&E-Stained Histological Images using Deep Learning Framework: A Perspective on Ensemble

Approach and Nuclei Count," 2023 IEEE 11th Region 10 Humanitarian Technology Conference (R10-HTC), Rajkot, India, 2023, pp. 462-467, doi: 10.1109/R10-HTC57504.2023.10461806.

[12] O. Berezsky, O. Pitsun, T. Datsko, I.Tsmots, V.Teslyuk. Specified diagnosis of breast cancer on the basis of immunogistochemical images analysis - Ceur Workshop Proceedings, 2020, 2753, pp. 129–135 https://ceur-ws.org/Vol-2753/short5.pdf

[13] R. Saha, M. Bajger and G. Lee, "Prior Guided Segmentation and Nuclei Feature Based Abnormality Detection in Cervical Cells," 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE), Athens, Greece, 2019, pp. 742-746, doi: 10.1109/BIBE.2019.00139

[14] A. B. Silva et al., "CNN Ensembles for Nuclei Segmentation on Histological Images of OED," 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS), L'Aquila, Italy, 2023, pp. 601-604, doi: 10.1109/CBMS58004.2023.00286

[15] A. B. Silva et al., "CNN Ensembles for Nuclei Instance Segmentation in OED Histological Images," 2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS), Madrid, Spain, 2025, pp. 369-374, doi: 10.1109/CBMS65348.2025.00082.

[16] P. Das, R. Sharma, S. Dey Roy, N. Nath and M. K. Bhowmik, "Ensemble Segmentation of Nucleus Regions from Histopathological Images towards Breast Abnormality Detection," 2022 25th International Conference on Computer and Information Technology (ICCIT), Cox's Bazar, Bangladesh, 2022, pp. 1137-1142, doi: 10.1109/ICCIT57492.2022.10055451.

[17] M. Kadaskar and N. Patil, "Nuclei Classification in Histopathology Images Using Fuzzy Ensemble of Convolutional Neural Networks," 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 2023, pp. 1-6, doi: 10.1109/ICCCNT56998.2023.10308315.

[18] O. Berezsky, O. Pitsun, N. Batryn, K. Berezska, L.Dubchak. Modern automated microscopy systems in oncology - Ceur Workshop Proceedings, 2018, 2255, pp. 311–325 https://ceur-ws.org/Vol-2255/paper28.pdf

[19] O. Berezsky, O. Pitsun, G. Melnyk, B. Derysh, P. Liashchynskyi. Application Of MLOps Practices For Biomedical Image Classification - Ceur Workshop ProceedingsOpen source preview, 2022, 3302, pp. 69–77 https://ceur-ws.org/Vol-3302/short3.pdf