Ensemble method for classifying students' stress factors*

Oleh Pitsun[1,*,†], Yaroslav Dykyi[1,†] , Maria Lysyk[1,†] ,Anastasiia Sobchak[1,†] and Maria Mudrak[1,†]

[1] *West Ukrainian National University, Ternopil 46003, Ukraine*

Abstract

The work is intended to solve the problems of classifying stress indicators among students using ensemble methods. Determining the level of stress and related indicators that affect the formation of stress allows you to identify factors that can affect the level of education and health of students. Existing algorithms for classifying quantitative indicators may not always show the best results, so this work uses the ensemble method to determine the best indicators. The research was conducted on the basis of the NEO-FFI-3 test and Student Adjustment to College Questionnaire (SACQ). Special attention is paid to ensemble learning methods, including Random Forest, Extra Trees, Gradient Boosting, Bagging and Voting Classifiers with hard and soft voting strategies. The experimental results demonstrate that ensemble-based models provide higher classification accuracy compared with individual classifiers such as k-Nearest Neighbors, Support Vector Machines, Logistic Regression and Naive Bayes. The research results are confirmed by a high classification quality indicator, namely more than 98 percent. The obtained results confirm the effectiveness of ensemble learning for analyzing complex multidimensional student stress indicators.

Keywords

Ensembles, Stress factors, Statistical data

1. Introduction

The educational process affects students in different ways in the world and in Ukraine in particular. The main factor that affects student success is stress, the level of satisfaction with learning. High levels of stress can negatively affect the psycho-emotional state of students, affect the level of learning, quality of sleep, the process of self-realization and the search for a future profession. Understanding the impact of various factors helps educational institutions improve learning conditions and, accordingly, improve the quality of students' knowledge. The human brain cannot fully analyze a large amount of data, find patterns and draw conclusions. To solve this problem, a large number of specialists or modern data analysis technologies are required. The use of modern methods of data analysis and classification allows you to determine patterns characteristic of a specific level of stress. The peculiarity is that it is necessary to find patterns in a large volume of data. The data is formed on the basis of a survey of students who fill out a questionnaire with more than 20 items.

Students of the Western Ukrainian National University in the number of more than 200 respondents participated in the study. The survey was formed on the basis of the NEO-FFI-3 and Student Adjustment to College Questionnaire (SACQ) Baker & Siryk methods. The questionnaire was formed on the basis of the evidence-based tests NEO-FFI-3 and Student Adjustment to College Questionnaire (SACQ). The approaches used allow us to determine the level and quality of

* *The Third International Conference of Young Scientists on Artificial Intelligence for Sustainable Development: Intelligent Systems and Artificial Intelligence (YAISD-WS 2026:ISAI) May 7-8, 2026, Ternopil, Ukraine*

[1,*] Corresponding author.

[†] These authors contributed equally.

o.pitsun@wunu.edu.ua (O. Pitsun); dykyyyaroslav@gmail.com (Y. Dykyi); marialysyk01@gmail.com (M. Lysyk); a.sob4aak@gmail.com (A. Sobchak); m.mudrak@wunu.edu.ua (M. Mudrak)

0000-0003-0280-8786 (O. Pitsun); 0009-0007-8510-0594 (Y. Dykyi); 0009-0000-2114-8272 (M. Lysyk); 0009-0001-8662-7001 (A. Sobchak); 0000-0003-3610-2490 (M. Mudrak)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-11

CEUR-WS.org/Vol-4326/paper16.pdf

learning, students' internal well-being, behavior. The influence of information factors and environmental factors.

To date, there are a large number of classification and analysis tools that allow us to analyze various quantitative characteristics in various areas of human life, including in determining stress factors. Standard approaches include factor analysis and regression modeling. They have certain limitations in dealing with large volumes of data. Machine learning methods, such as tree-like algorithms, random forests, gradient boosting, and neural networks, provide more accurate results by identifying hidden relationships in the data. The ensemble method, which combines several models to improve the accuracy of prediction, becomes especially relevant for the tasks of classifying complex phenomena, such as stress factors among students.

Object of study - data on the level of stress among students according to the results of the study based on known approaches.

The subject of the study is ensemble methods for data classification.

The purpose of this article is to analyze stress indicators among students by using modern approaches to the classification of quantitative indicators based on the ensemble method, which allows to increase the accuracy of classification.

The novelty of this study lies in the development of a unified ensemble-learning framework for the classification of student stress factors. Unlike traditional approaches that rely on a single machine learning algorithm, the proposed framework combines multiple ensemble strategies.

2. Literature review

In [1], the study aimed to analyze the psychological state of students and identify the relationship between stress levels and time spent on the Internet. The data set collected at Jaypee Institute of Information Technology, which contained information on 206 students, was used for the analysis. Four classification algorithms were used: linear regression, naive Bayesian classifier, random forest, and support vector machine (SVM). The results were evaluated using indicators such as sensitivity, specificity, and accuracy. The results were then analyzed to assess the effectiveness and accuracy of the models.

In [2], a study is presented that can help teachers and administrators of educational institutions more effectively direct resources to reduce stress levels among female engineering students.

In [3], a hybrid model combining various computational methods and using both supervised learning (SL) and stochastic methods (SC) approaches is presented, which can solve research and exploitation problems through adaptation, parameter tuning, and the use of chaos, Lévy distribution, and normal (Gaussian) distribution.

To calculate the student stress index, in [4], a two-stage Temporal Dynamic Bayesian Network (TDBN) model was created. This model determines the level of stress based on four parameters: evidence from leaf nodes, study load, context, and individual health characteristics of the student.

In a study [5] on the prevalence of stress, 355 students from 28 different universities in Bangladesh were surveyed. The questionnaire included questions on anthropometric indicators, academic aspects, lifestyle, and health data, which allowed to assess the subjective perception of stress (yes or no). The Boruta algorithm was used to identify significant predictive factors of stress occurrence.

The method proposed in [6] uses a set of stimuli to analyze the relationship between stress levels and increased cognitive load. After the experiment is completed, an eye strain test is performed. A questionnaire is completed at the end of the study to assess the student’s comfort zone.

In study [7], the influence of external factors on stress levels was analyzed. The results of the study show that watching films that promote positive thinking helps to manage stress and increases the level of creative problem solving in students in all categories, which is confirmed by pre- and post-test data.

In study [8], 115 students with a mean age of 19±6.76 years participated. In total, 35 students (30.4%) had minor or moderate physical problems, 20 (17.4%) had serious physical problems, and 60

students (52.2%) had no physical problems. The mean stress score was 19.6±6.76. The main factors contributing to stress were environmental factors, new learning environment, bullying of students, challenging curricula and personal factors.
The aim of the study [9] was to evaluate the model of vulnerability to stress among French university students. Stress factors were assessed using six scales that were available online for three months. The study involved 483 students aged 18 to 24 years (mean = 20.23, standard deviation = 1.99). The results showed that 72.9%, 86.3% and 79.3% of them suffered from psychological distress, anxiety and depressive symptoms, respectively.
The work [10-12] presents approaches to data analysis and processing using artificial intelligence.

3. Problem statement

To conduct the research, the following tasks need to be implemented:
- develop a questionnaire structure based on generally accepted approaches;
- analyze data using visual graphical analysis;
- develop and implement an ensemble method for classifying the obtained data.

4. Dataset

The developed dataset was formed based on a survey of more than 200 students of the West Ukrainian University. The developed dataset "WUNUTH-24" was formed based on a survey of students of technical and humanitarian specialties. An example of the data is shown in Figure 1.

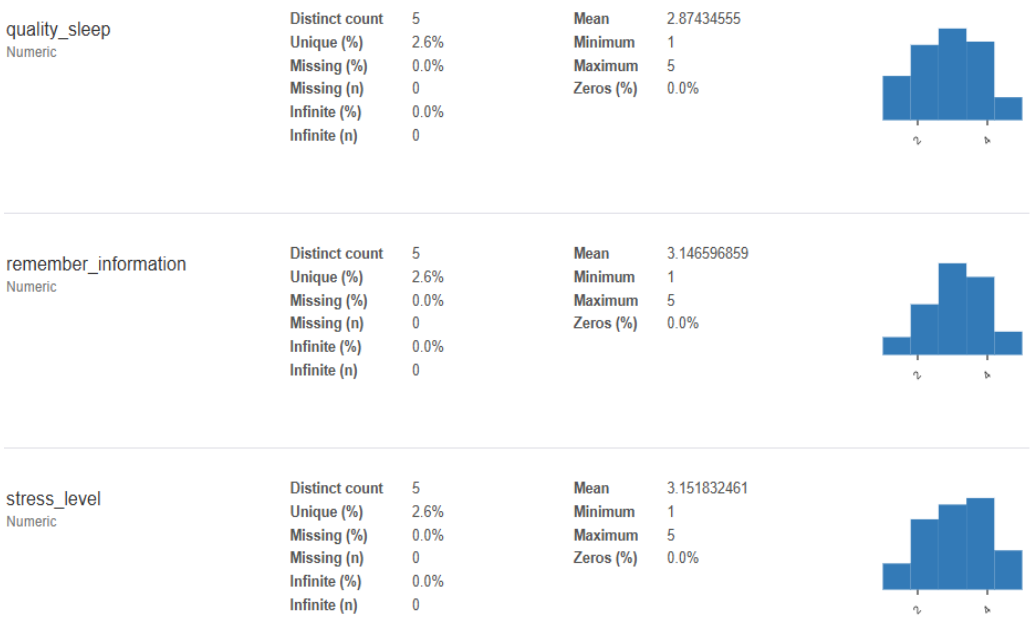


Figure 1: Record distribution.

This figure shows only a portion of the proposed dataset. The use of students from different majors is designed to avoid situations related to the specificity of learning.

5. Materials and research methods

Unlike traditional approaches that rely on a single classifier, the proposed framework integrates multiple ensemble learning strategies within a unified processing pipeline. Proposed framework enabling improved classification robustness and higher predictive accuracy for student stress assessment.

The process starts with the data collection stage. Data are obtained from the NEO-FFI-3 questionnaire, the Student Adjustment to College Questionnaire (SACQ). These data describe students' psychological and behavioral characteristics.

The next stage is data preprocessing. Missing values are identified and processed. Incorrect or inconsistent records are removed. Categorical variables are converted into numerical form.

After preprocessing, exploratory data analysis is performed. Correlation analysis is used to identify relationships between variables. Data visualization helps reveal important patterns and trends.

The normalized data are used for baseline model training. Several machine learning algorithms are tested, including Logistic Regression, Support Vector Machine, Decision Tree, and k-Nearest Neighbors. These models provide reference results for comparison.

The next stage is ensemble model construction. Different ensemble techniques are applied, including bagging, boosting, and voting methods.

The developed models undergo model evaluation, where their predictive performance is assessed using classification metrics such as accuracy, precision, recall and F1-score. The evaluation results are then used to identify the most effective approach for classifying student stress factors.

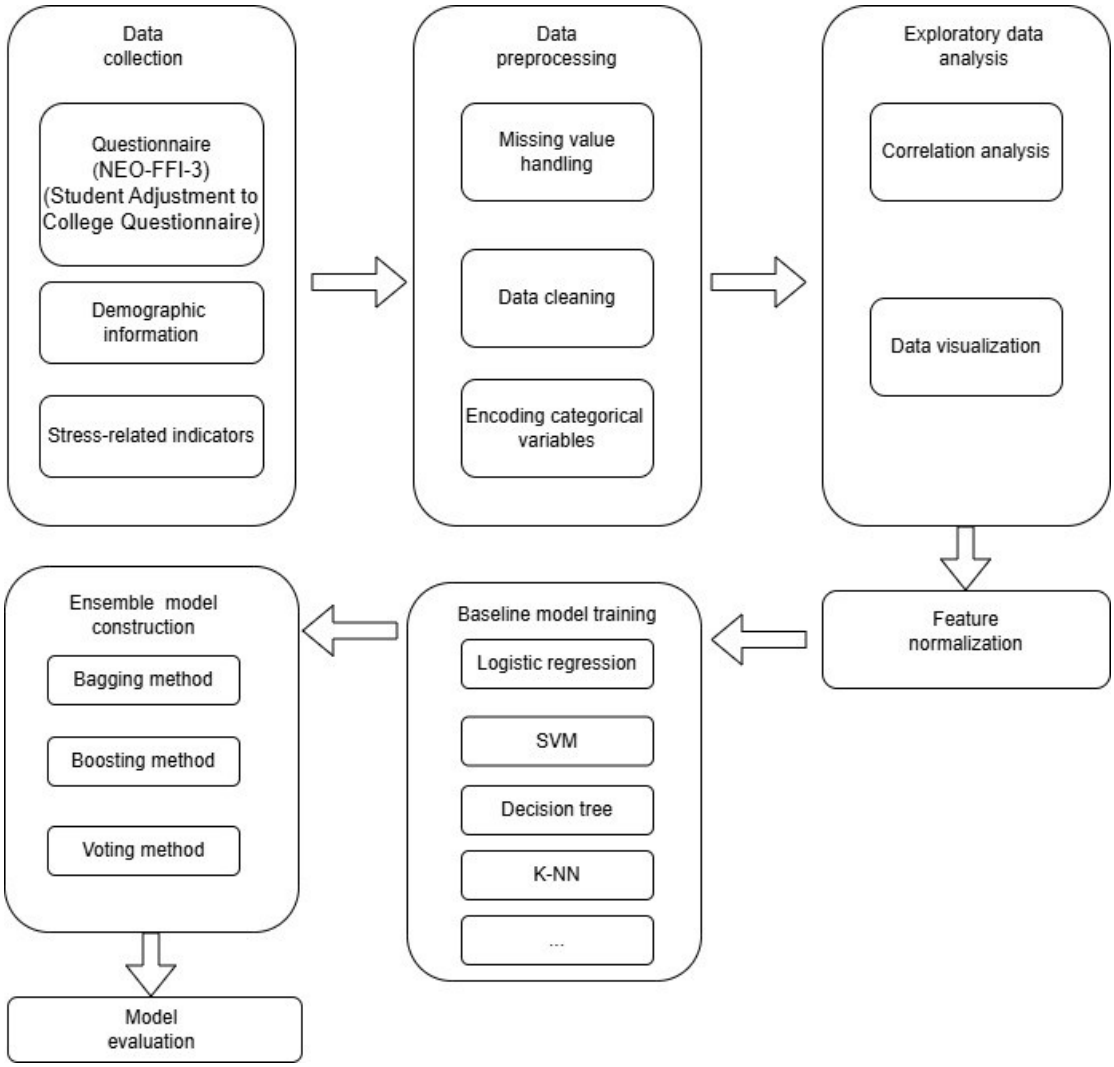


Figure 2: Proposed framework

Compared with hard voting, the soft voting strategy often provides improved classification accuracy and better robustness in the presence of data noise. In addition, it can be implemented using different types of base classifiers.

$$y = argmax \sum_{j=1}^{m} w_j p_{ij} \quad (1)$$

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

The support vector machine (SVM) method builds a hyperplane that maximally separates classes using support vectors. This is a specialized variant of SVM for linearly separable data, which uses an optimized decision algorithm (liblinear). It works faster than a regular SVM and is suitable for large samples. It uses L2 regularization, which helps to avoid overfitting. The main C parameter determines the rigidity of the separation boundaries between classes. The algorithm is less efficient for complex nonlinear boundaries.

k-NN classifies a new object by its k nearest neighbors in the feature space. It does not build a model in advance, but memorizes all the data, so it requires a lot of memory. The algorithm is sensitive to the choice of k: a small value can lead to noisy results, and too large - to loss of detail. GridSearchCV helps to choose the optimal value of k and the distance metric.

Decision trees are trained by sequentially dividing the feature space according to gini or entropy criteria. They are easy to interpret and can work with both quantitative and categorical features. The main disadvantage is the tendency to overtraining if the tree is too deep. Using the max_depth parameter helps control this.

An ensemble method that combines predictions from multiple base models using majority voting. Hard voting means that each model votes for a specific class, and the class with the most votes is selected. It improves stability and generalization ability if individual models are weakly correlated. It is easy to implement and effective for reducing random errors. However, it does not work as well if individual classifiers have very different performance levels.

Soft voting takes into account the probabilities of each model's predictions, rather than simply counting votes. This allows more influential models to have a greater influence on the final decision. Using GridSearchCV allows you to find the optimal combination of models and their weights. The method works well when the models complement each other, reducing the overall error. However, it is sensitive to models that overestimate or underestimate probabilities.

6. Computer experiments

The proposed framework provides a systematic workflow for transforming questionnaire-based student data into reliable predictive models and demonstrates the potential of ensemble learning techniques for educational and psychological data analysis.

Fig. 3 shows a matrix of correlations between the studied indicators: df_index, state_during_last_month, headaches, level_of_concentration, quality_food, quality_sleep, remember_information and stress_level. The values of the correlation coefficients vary in the range from -1 to 1. Red color reflects a positive relationship, and blue is a negative one.

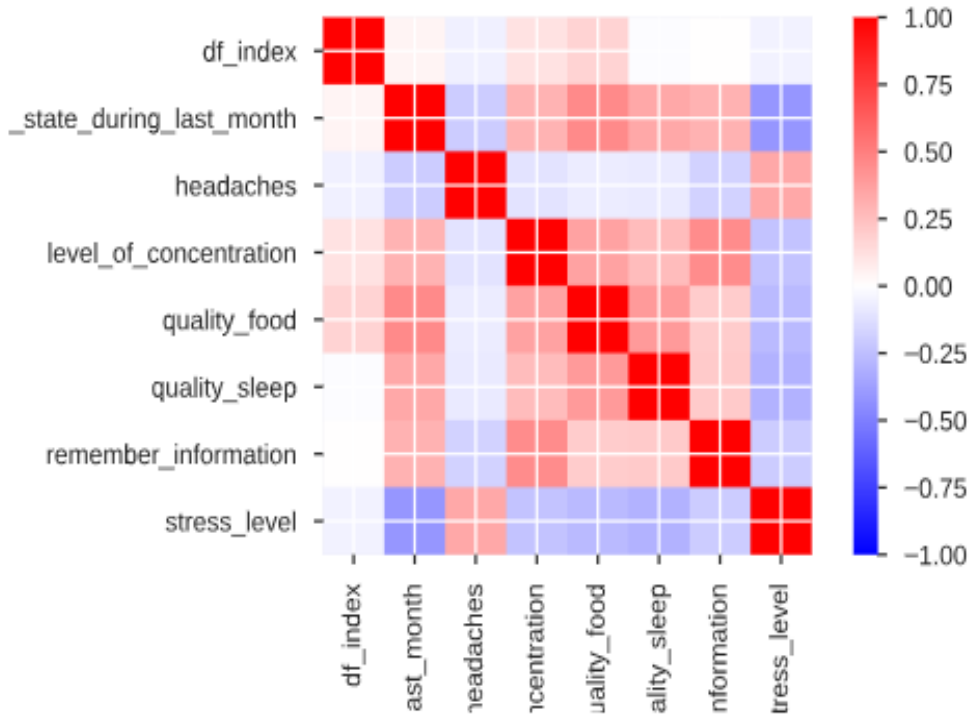


Figure 3: Correlation matrix.

A correlation matrix was constructed to analyze relationships between the collected variables. The purpose of this analysis was to identify significant positive and negative correlations, explore dependencies among stress-related factors, and gain a better understanding of the dataset before training machine learning models.

The correlation matrix was visualized using a heatmap, allowing for a clear representation of the strength and direction of relationships between the analyzed features.

Matrix analysis shows that the most pronounced positive correlations are observed between the indicators quality_food, quality_sleep, level_of_concentration and remember_information. The stress_level indicator shows predominantly negative correlations with most variables. The strongest inverse relationship is observed between stress level and sleep quality, attention span, and the ability to remember information. This indicates that increasing stress is accompanied by a deterioration in cognitive functions and general well-being. The variable headaches also has a negative relationship with indicators of concentration, memory, and sleep quality, although the strength of these correlations is moderate.

The study has several limitations. First, the dataset was collected from a single university. Second, the analysis is based on self-reported questionnaire responses.

The structure of the neural network is shown in Figure 3.

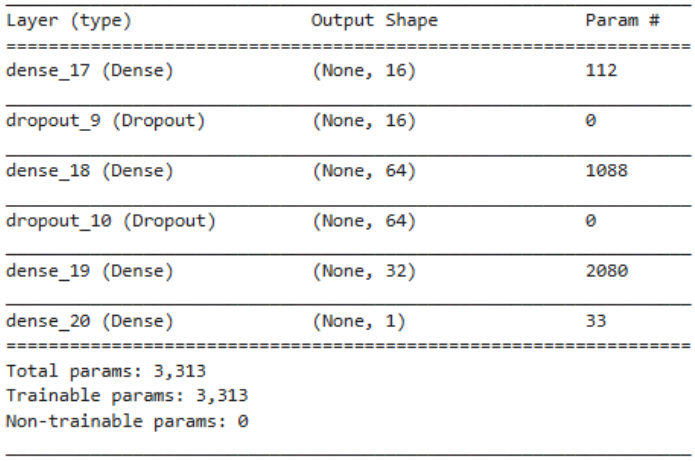


Figure 4: Neural network structure.

The total number of network parameters is 3313, all of which are trainable (Trainable parameters = 3313), and there are no untrainable parameters. This indicates a compact architecture and low computational complexity compared to deep neural networks. A detailed analysis of the results is given in Table 1.

Table 1.
Analysis of the results obtained.

Model	Score
VotingClassifier-soft voting	98.03
Random Forest	98.03
GradientBoostingClassifier	86
ExtraTreesClassifier	98
Decision Tree Classifier	98
VotingClassifier-hard voiting	97.37
BaggingClassifier	94.08
k-Nearest Neighbors	68.42
Support Vector Machines	63.82
XGBClassifier	49.34
Logistic Regression	49.34
LGBMClassifier	44.08
AdaBoostClassifier	43.42
Linear SVC	42.76
RidgeClassifier	42.11
Naive Bayes	40

Stochastic Gradient Decent	38.16
Perceptron	25.66

Analyzing the research results, we can conclude that the VotingClassifier-soft voting algorithm showed the best results, which allowed us to determine it as a priority when choosing algorithms for classifying quantitative indicators of the psychological state of students.

Conclusion

In this work, a mechanism for classifying quantitative characteristics of indicators of the psychological state of students of the Western Ukrainian National University has been developed. A comparative analysis of the use of ensemble methods allows us to establish that the VotingClassifier-soft algorithm showed the best result and is above 98% by the accuracy criterion. The obtained results indicate that combining predictions from multiple complementary models improves robustness, reduces variance and increases the generalization ability of the classification system. An important contribution of this research is the development of an integrated ensemble-learning framework. This proposed framework combines data preprocessing, exploratory data analysis, feature normalization, ensemble model generation, performance evaluation within a unified workflow. Future research will focus on improving ensemble models by incorporating additional combination strategies. Explainable AI methods, such as SHAP and LIME, will be used to better understand the influence of different stress factors. Further work will also include the development of real-time stress monitoring systems for continuous assessment of student well-being.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] R. WanAhuja, A. Banga, Mental stress detection in university students using machine learning algorithms, Procedia Computer Science 152 (2019) 349–353. doi:10.1016/j.procs.2019.05.007.

[2] A. Rizwan, S. Farooq, M. S. I. Alvi, S. Nawaz, Analysis of factors affecting the stress level of female engineering students, Global Journal of Human-Social Science 12 (2012).

[3] S. Sharma, G. Singh, M. Sharma, A comprehensive review and analysis of supervised-learning and soft computing techniques for stress diagnosis in humans, Computers in Biology and Medicine 134 (2021) 104450. doi:10.1016/j.compbiomed.2021.104450.

[4] P. Verma, S. K. Sood, A comprehensive framework for student stress monitoring in fog-cloud IoT environment: m-health perspective, Medical & Biological Engineering & Computing 57 (2019) 231–244. doi:10.1007/s11517-018-1877-1.

[5] R. Rois, M. Ray, A. Rahman et al., Prevalence and predicting factors of perceived stress among Bangladeshi university students using machine learning algorithms, Journal of Health, Population and Nutrition 40 (2021) 50. doi:10.1186/s41043-021-00276-5.

[6] C. Jyotsna, J. Amudha, Eye Gaze as an Indicator for Stress Level Analysis in Students, in: 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, India, 2018, pp. 1588–1593. doi:10.1109/ICACCI.2018.8554715.

[7] R. Rastogi, D. K. Chaturvedi, S. Satya, N. Arora, P. Singh, P. Vyas, Statistical analysis for effect of positive thinking on stress management and creative problem solving for adolescents, in: Proceedings of the 12th INDIACom, 2018.

[8] K. Qamar, N. S. Khan, M. R. Bashir Kiani, Factors associated with stress among medical students, Journal of the Pakistan Medical Association 65(7) (2015) 753–755. PMID: 26160086.
[9] D. Saleh, N. Camart, L. Romo, Predictors of Stress in College Students, Frontiers in Psychology 8 (2017) 19. doi:10.3389/fpsyg.2017.00019.
[10] O. Berezsky, O. Pitsun, G. Melnyk, V. Koval, Y. Batko, Multi-threaded Parallelization of Automatic Immunohistochemical Image Segmentation, in: Z. Hu, Y. Wang, M. He (Eds.), Advances in Intelligent Systems, Computer Science and Digital Economics IV. CSDEIS 2022, Lecture Notes on Data Engineering and Communications Technologies, vol. 158, Springer, Cham, 2023. doi:10.1007/978-3-031-24475-9_23.
[11] O. Berezsky, P. Liashchynskyi, O. Pitsun, P. Liashchynskyi, M. Berezkyy, Comparison of Deep Neural Network Learning Algorithms for Biomedical Image Processing, CEUR Workshop Proceedings 3302 (2022) 135–145. URL: https://ceur-ws.org/Vol-3302/paper7.pdf.
[12] O. Pitsun, O. Berezsky, G. Melnyk, Y. Batko, P. Liashchynskyi, M. Berezkyi, MLOps Approach for Automatic Segmentation of Biomedical Images, CEUR Workshop Proceedings (2023) 241–248. URL: https://ceur-ws.org/Vol-3302/paper7.pdf.