

## An approach toward improvement of ensemble method's accuracy for biomedical data classification

Ivan Izonin<sup>1,2</sup>, Roman Muzyka<sup>2</sup>, Roman Tkachenko<sup>3</sup>, Michal Gregus<sup>4</sup>, Natalya Kustra<sup>3</sup>,  
Stergios-Aristoteles Mitoulis<sup>1</sup>

<sup>1</sup>Department of Civil Engineering, School of Engineering, University of Birmingham, Birmingham, United Kingdom

<sup>2</sup>Department of Artificial Intelligence, Lviv Polytechnic National University, Lviv, Ukraine

<sup>3</sup>Department of Publishing Information Technologies, Lviv Polytechnic National University, Lviv, Ukraine

<sup>4</sup>Faculty of Management, Comenius University Bratislava, Bratislava, Slovak Republic

### Article Info

#### Article history:

Received Mar 9, 2024

Revised Jul 9, 2024

Accepted Jul 17, 2024

#### Keywords:

Biomedical engineering

Classification task

Ensemble method

Imbalanced dataset

Improved accuracy

Ito decomposition

Machine learning

### ABSTRACT

Amidst rapid technological and healthcare advancements, biomedical data classification using machine learning (ML) is pivotal for revolutionizing medical diagnosis, treatment, and research by organizing vast healthcare-related data. Despite efforts to apply single ML models on clean datasets, satisfactory classification accuracy can still be elusive. In such cases, ML-based ensembles offer a promising solution. This paper explores cascaded ensembles as highly accurate methods. Existing cascade classifiers often partition large datasets into equal unique parts, limiting accuracy due to insufficient amount of useful information processed by weak classifiers of all levels of the cascade ensemble. To address this, we propose an improved cascaded ensemble scheme using a different data sampling approach. Our method forms larger subsamples at each cascade level, enhancing accuracy, and generalization properties during biomedical data analysis. Experimental comparisons demonstrate substantial increases in classification accuracy and generalization properties of the improved cascade ensemble.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



### Corresponding Author:

Roman Muzyka

Department of Artificial Intelligence, Lviv Polytechnic National University

S. Bandera str., 12, Lviv, 79013, Ukraine

Email: roman.muzyka.mknssh.2022@lpnu.ua

## 1. INTRODUCTION

In the 21st century, amidst the rapid advancements in technology and healthcare, the significance of biomedical data classification in medicine cannot be overstated. Biomedical data classification, the process of organizing and categorizing vast amounts of healthcare-related data, plays a pivotal role in transforming the landscape of medical diagnosis, treatment, and research [1], [2]. First of all, biomedical data classification serves as the cornerstone of accurate and efficient disease diagnosis. In the era of precision medicine, where treatments are tailored to individual patients, the ability to analyze diverse datasets—from medical imaging and genetic information to clinical parameters—is paramount. Classification algorithms based on machine learning (ML) or artificial neural networks (ANN) can sift through this wealth of information, identifying patterns and signatures indicative of specific diseases or conditions [3]. By leveraging ML or ANN, clinicians can make more informed diagnostic decisions, leading to earlier detection and intervention.

Biomedical data classification holds immense promise in the realm of personalized medicine [4]. Each patient is unique, with distinct genetic backgrounds, lifestyle factors, and treatment responses [5]. Classification models can analyze individual patient data, including genomic profiles, biomarkers, and treatment histories, to predict personalized treatment outcomes [6]. By identifying the most effective

therapies for each patient while minimizing adverse effects, personalized medicine promises to revolutionize healthcare delivery, ushering in an era of truly patient-centered care.

Also, biomedical data classification empowers clinicians with invaluable decision support tools. In today's complex healthcare landscape, where medical knowledge is expanding exponentially, clinicians are inundated with vast amounts of data and information [7]. Classification algorithms can serve as clinical decision support systems, synthesizing patient data with evidence-based guidelines and best practices to provide actionable insights. By offering recommendations for diagnosis, treatment selection, and patient monitoring, these systems enhance the quality and efficiency of healthcare delivery, ultimately improving patient outcomes.

In addition to its clinical applications, biomedical data classification fuels groundbreaking discoveries in medical research [8]. The advent of high-throughput technologies, such as next-generation sequencing and omics platforms, has generated unprecedented volumes of data, providing researchers with unprecedented insights into human health and disease. Classification algorithms can analyze these large-scale datasets, identifying novel disease biomarkers, elucidating disease mechanisms, and predicting treatment responses [9]. By accelerating translational research and the development of innovative therapeutics, biomedical data classification drives progress toward improved treatments and cures for a myriad of diseases.

Furthermore, biomedical data classification plays a crucial role in public health surveillance and epidemiology [10]. By analyzing population-level data, including demographic trends, environmental factors, and disease prevalence rates, classification models can detect disease outbreaks, identify high-risk populations, and inform public health interventions. From infectious disease surveillance to chronic disease management, biomedical data classification serves as a powerful tool for safeguarding public health and promoting well-being on a global scale [11]. However, all these domains necessitate precise solutions when scrutinizing extensive biomedical datasets. Furthermore, with the ongoing evolution of technology leading to the generation of progressively intricate datasets, a multitude of challenges emerges, particularly concerning the precision of their analysis through prevailing methods and models.

A thorough review of the scientific literature has allowed researchers to identify several strategies used to improve biomedical dataset's classification accuracy. The first arises from the peculiarities of biomedical datasets, namely a large number of features upon which the classification process is based [12]. Some of these features may be irrelevant or redundant. Feature selection/extraction, and feature importance techniques [13] can be employed to reduce dimensionality and improve model performance [14]. This approach ensures the removal of uninformative or weakly informative features, which decrease prediction accuracy. Additionally, it allows for the reduction of the input data space, enhancing the generalization properties of ML or ANN underlying the data classification process. Moreover, this approach reduces the duration of training procedures for selected methods, as they work with a significantly smaller number of features compared to the initial dataset [15]. Despite the significant advantages of such an approach, the use of similar methods may reduce the amount of relevant information [16]. Removing certain features from the dataset may lead to the elimination of important patterns or relationships that could contribute to better model performance. Furthermore, feature selection methods may inadvertently lead to overfitting, where the model performs well on the training data but fails to generalize to unseen data. This can occur if the feature selection process is overly tailored to the training dataset and does not consider the variability present in the broader population.

Another important technique is the incorporation of domain-specific knowledge into feature engineering or model design [17]. Biomedical experts can provide valuable insights into relevant features or relationships within the data, which can guide the model-building process [18]. However, such an approach requires substantial knowledge in the domain area of the data, which may not always be accessible to machine learning engineers.

In the case of analyzing an imbalanced dataset, which is quite typical for biomedical data, a crucial technique for ensuring the reliability of the chosen ML or ANN and improving the accuracy of classification results is data balancing [19]. Techniques such as oversampling, undersampling, or using class weights can help mitigate this issue and improve the model's ability to generalize across classes [20]. However, in this case, it is important to carefully select the necessary technique, conduct a series of experimental studies on its effectiveness, and only then apply it in practice.

In case of unsatisfactory performance accuracy of ML or ANN on an already prepared dataset, for which all the techniques mentioned above have been qualitatively executed, it is necessary to correctly select the ML or ANN [21]. Today, there are a plethora of different methods, models, or machine learning algorithms, built on various principles and designed to solve tasks of a certain class. Selecting the right method will ensure high classification accuracy with satisfactory time resources for its operation. Another equally important element of applying such an approach is hyperparameter tuning [22]. Grid search, randomized search or some other optimization models can be employed to systematically search for optimal hyperparameters [23]. The productivity of the entire method largely depends on this step.

However, cases often arise when correctly selected single models with optimal parameters operating on cleaned datasets still do not provide satisfactory classification results [24]. In such cases, ensemble methods can be applied, which combine different or identical ML or ANN methods, referred to as ensembles [25]. Among the four most well-known classes of ensembles are bagging, boosting, stacking, and cascading. These techniques can further enhance predictive performance by leveraging diverse or identical models or by focusing on instances that are challenging to classify.

Among the problems faced by all four classes of ensemble methods is the selection of the ML or ANN methods that will form the basis of a particular ensemble [26]. Experimenting with different base learners can yield better performance as each base learner may capture different aspects of the data and contribute differently to the ensemble's performance [27]. Speaking of heterogeneous stacking ensembles or bagging, ensemble diversity should be considered here. Diversity can be achieved by using different algorithms, different subsets of features, or training the models on different portions of the dataset [28]. Another characteristic shared by all four classes of ensemble methods is the size or depth (for cascading) of the ensemble. Increasing the ensemble size or depth can improve performance up to a certain point, after which further additions may yield diminishing returns or increased computational cost. All of this should be taken into account when composing a particular ensemble method for solving classification tasks with biomedical datasets.

In this article, we examine cascading ensembles as one of the most accurate classes of these methods [29], [30]. In addition to high accuracy, cascading ensembles almost always mitigate overfitting issues and demonstrate high generalization properties. A particular feature of these models is that base models are trained hierarchically. Each model is trained using the outputs of all preceding models in the hierarchy [31]. For instance, in [32], the problem of human activity recognition is considered based on the use of a cascading classifier. The authors developed a multilevel cascading scheme, where at each level, four different machine learning algorithms are employed, returning probabilities of belonging to each of the defined task classes. These probabilities are combined with the initial inputs of the task and serve as input features for the next level of the cascade. This approach provided a significant increase in classification accuracy, although it significantly expands the input data space of each cascade level, especially when constructing a multilevel scheme. In [33], another cascading scheme is developed. The authors used only one machine learning algorithm, which is used at each new level of the cascade. The classification results of the lower-level algorithm are transmitted as an additional feature for the higher-level classifier [34]. Additionally, this cascade differs from others due to additional modeling of relationships between input attributes and the output of the lower-level algorithm. Such a relationship is modeled using Ito decomposition. In this case, the input data space, as in [32], is also expanded, but Ito decomposition ensures higher prediction accuracy. Furthermore, the advantage of this cascade is the use of only one ML at each level of the cascade. In case [32], this is a linear method, the use of which is justified in terms of the duration of the training procedure for the entire cascade construction.

In general, the use of linear, i.e., high-speed artificial intelligence (AI) tools during the analysis of large datasets is justified in terms of performance [35]. This also applies to ensemble methods, including cascading implementations of these methods. However, most medical tasks require high accuracy, which is difficult to achieve using such methods. Therefore, the cascading scheme from [32] modeled non-linearity by using Ito decomposition. Additionally, a particular feature of this cascade, as well as many others, is that the input dataset of large volumes is divided into levels or nearly equal parts. The number of parts determines the number of levels in the cascade. However, in the case of a cascading scheme with a large number of levels, the classifier at each level of the cascade receives only a small portion of data for analysis. Such incomplete information at each level of the cascade does not ensure obtaining the maximum possible accuracy that cascading ensembles can provide [36].

Therefore, this paper aims to improve the accuracy of classifying large volumes of biomedical data based on modifying the cascading ensemble of linear artificial intelligence tools by applying a new data sampling scheme for implementing training procedures at each level of the cascade. The main contribution of this paper is the following:

- We have enhanced the existing cascade ensemble by employing a new data sampling scheme to implement artificial intelligence training procedures at each new level of the ensemble. Compared to the existing scheme of uniform data division for each level of the cascading ensemble, the developed scheme differs in that for each level of the cascade, a significantly larger subsample (a certain percentage of the total dataset) is randomly formed, the size of which is determined by the user. Essentially, subsamples with replacements are formed here. Thus, weak classifiers at each level of the cascade can receive a significantly larger amount of data for training procedures, ensuring a significant improvement in the accuracy of the cascading ensemble when analyzing biomedical datasets.
- We conducted an experimental comparison of the accuracy results between the existing and modified cascade ensembles in this study and found that the modified cascade ensemble significantly improves classification accuracy while enhancing the method's generalization properties overall.

The paper is structured as follows: in section 2, we delve into the theoretical framework underpinning our research, providing a nuanced understanding of the concepts that inform our methodologies. Section 3 presents detailed modeling information, including optimal parameter selection and the results obtained. Additionally, this section includes a comparative analysis with existing methods to contextualize our findings within the broader scholarly discourse. Finally, in the last section, we propose avenues for further research to build upon our discoveries and address any lingering questions that arise from this study.

## 2. METHOD

As mentioned above, in [32], a cascade ensemble based on linear machine learning methods was developed for analyzing biomedical datasets. The original article addressed a regression problem, so in this work, we adapted this method for solving a classification task. Linear methods form the basis of the existing ensemble from [32] because they provide high-speed operation. This is one of the critical limitations of intelligent systems for analyzing large datasets. However, such methods often do not provide sufficient prediction accuracy. To address this issue, the cascade ensemble from [32] used Ito decomposition at each cascade level. This approach allowed accounting for nonlinearities within the given dataset. Furthermore, it provided additional modeling of relationships between the features of the dataset. Additionally, considering the composition of the cascade ensemble, where the output of the previous cascade level is used as an additional attribute in the next one, Ito decomposition also modeled relationships between all inputs and the predicted output, further linearizing the response surface.

Taking all of the above into account, the existing cascade ensemble provided a significant improvement in prediction accuracy while maintaining the satisfactory computational resources required for its operation. However, a particular feature of this method is that the given training dataset was randomly distributed into unique subsamples of equal size, which were processed at each level of the cascade. It should be noted that their quantity determined the number of levels in the cascade ensemble from [32]. However, in the case of a dataset of moderate size or the construction of a multi-level cascade ensemble, a weak classifier at each level would receive significantly less useful information for correct classification. This would affect the accuracy of the entire method. Therefore, in this paper, we proposed a new scheme for dividing the training set into subsamples. According to it, the number of levels in the cascade construction defined by the user will determine the number of necessary subsamples for operation at each level. In our case, such subsamples will also be formed randomly based on the training dataset. However, here, it will not be an equal division but the possibility of forming a subsample with the volume determined by the user (for example, each of the  $M$  subsamples will contain 60% of vectors randomly selected from the training dataset). Thus, the new scheme will involve forming subsamples with replacements containing significantly more data for processing by classifiers at each level of the cascade ensemble. This approach should increase the accuracy of the entire ensemble construction.

Since this work proposes a new scheme for dividing the dataset for the cascade ensemble, the basic procedures of its training and application remain unchanged. They are described in detail in [32]. However, let us briefly consider the training and application algorithms of the enhanced cascade ensemble. The training procedure according to the method involves the following steps.

- a. The user defines two initial parameters: the number of cascade levels, denoted as  $M$ , and the percentage of randomly selected vectors used to create  $M$  subsamples with repetitions from the total training dataset. This process involves forming subsamples with repetitions.
- b. The first subset undergoes processing using Ito decomposition, followed by training the classifier of the first level on the modified subset 1.
- c. The second subset undergoes processing using Ito decomposition as well. Subsequently, it is applied to the previously trained classifier of the first level of the cascade ensemble. The predicted value obtained is then added as an additional independent attribute. Additionally, the enlarged second subset, now including one additional attribute, undergoes processing using Ito decomposition. Finally, the classifier of the second level is trained on the modified subset 2.
- d. Steps (a-b) are repeated iteratively until only the last  $M$  subset remains.
- e. The final step involves sequentially traversing through the last  $M$  levels of the cascade. The output value of the last level of the cascade ensemble will represent the sought attribute.

The flowchart illustrating the enhanced cascade ensemble in its training mode for biomedical data classification is depicted in Figure 1. This flowchart outlines the sequential steps involved in training the cascade ensemble to optimize its performance in classifying biomedical data. In practice, deploying the pre-trained cascade ensemble for classifying new observations with unknown output attributes follows a specific procedure. Each new observation progresses through all levels of the cascade ensemble. At each level, the

ensemble refines its classification decision based on the available data and model parameters. This iterative process continues until the final level of the cascade, where the sought value of the class label, indicating the classification of the current observation, is determined. This approach ensures that each level of the cascade ensemble contributes progressively to the classification process, leveraging the strengths of each classifier within the ensemble to achieve accurate and reliable classification results. By structuring the classification process in this manner, the enhanced cascade ensemble enhances both efficiency and accuracy in biomedical data classification tasks, making it a robust methodology for handling complex and diverse datasets.

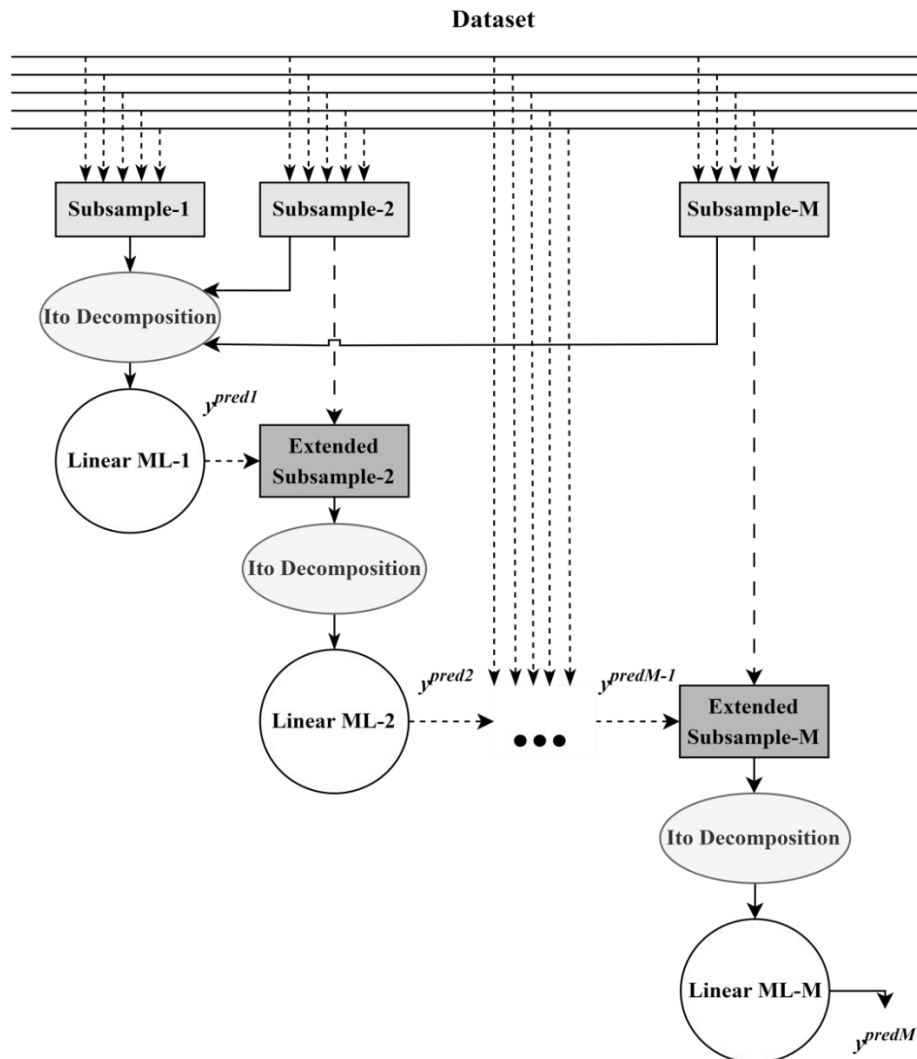


Figure 1. Flowchart of the improved cascade ensemble in training mode for biomedical data classification through the use of a new training data sampling scheme

### 3. RESULTS AND DISCUSSION

This section describes the dataset used for modeling and its preprocessing. The results of selecting the optimal parameters for the enhanced cascade ensemble are presented, and the obtained results are summarized based on various performance indicators. The authors developed their own software for implementing the enhanced cascade ensemble. Stochastic gradient descent (SGD) algorithm was chosen as the weak classifier for the cascade ensemble, characterized by high speed during the analysis of large biomedical datasets. Modeling was conducted on a computer with the following parameters:

- Processor: Intel® Core™ i7-8750H with a clock speed of 2.20 GHz
- Amount of RAM: 8 gigabytes
- Operating system: Microsoft Windows 10 Home.

### 3.1. Dataset descriptions and preprocessing

In this study, authors utilized publicly available data from the behavioral risk factor surveillance system (BRFSS) in the United States for the year 2021, sourced from the Centers for Disease Control and Prevention [37]. The initial dataset encompassed a broad spectrum of information on various diseases, prompting the authors to curate a final dataset focusing on lifestyle factors potentially associated with cardiovascular disease risk. The resulting dataset comprised 308,854 records with 29 attributes. During data processing, a notable imbalance emerged within the dataset, with a distribution of 92% to 8% across both task classes. To address this, the authors employed a data-balancing strategy leveraging both oversampling and undersampling techniques. Specifically, they utilized synthetic minority over-sampling technique (SMOTE) to augment instances of the smaller class and NearMiss to reduce instances of the larger class [38]. This approach was executed iteratively, adjusting parameters to vary the number of instances for each class within the dataset, ranging from 50,000 to 150,000 with increments of 25,000. The outcomes of this hybrid oversampling and undersampling methodology showcased strong generalization capabilities and satisfactory accuracy of the SGD classifier, particularly when balancing the data in a 75,000 to 75,000 ratio. Subsequently, this balanced dataset served as the training sample for subsequent experimental investigations.

### 3.2. Parameters tuning and results

The enhanced cascade ensemble, like the baseline method, is characterized by several parameters. Firstly, there is the number of levels in the cascade ensemble, which ensures the highest accuracy of its operation. However, in addition to the aforementioned parameter common to both cascade ensembles, the enhanced cascade ensemble is also characterized by another one—the percentage of vector utilization from the training sample in each new subset, randomly formed for each weak classifier of the cascade. This parameter arises due to the new data sampling scheme proposed in this article for implementing cascade training procedures. It affects both the method's accuracy and the duration of the training procedure. Both parameters are important when analyzing large biomedical datasets. Therefore, in this work, a detailed selection of both of these parameters was carried out to achieve the highest performance of the enhanced cascade ensemble.

Figure 2 presents the results of the study of the accuracy of the enhanced cascade ensemble (based on F1-score) when changing both the percentage of utilization of the training sample for forming new subsets with repetition for each level of the method and the number of such cascade ensemble levels. It is worth noting that the results are presented for both the training mode in Figure 2(a) and the application mode in Figure 2(b) to evaluate not only the accuracy of the method but also its generalization properties, which are extremely important during the practical use of the method.

Summarizing the results presented in Figure 2, the following conclusions can be made:

- Increasing the number of vectors in the subsamples with repetition for each level of the method slightly reduces the method's accuracy in the training mode but increases accuracy in the application mode. However, this occurs up to a certain point (70% of the training dataset volume). This holds for all four variants of the cascade ensemble, each differing in the number of levels (from 2 to 5).
- Using both small (20-30% of the training dataset volume) and large numbers of vectors (more than 70% of the training dataset volume) for forming subsamples with repetition demonstrates a deterioration in the method's generalization properties. The gap between training and application accuracy significantly increases.
- The highest generalization properties, regardless of the subsample size for each level of the method, are demonstrated by the three-level cascade ensemble. For ease of understanding, these results are separately presented in Figure 3.
- In terms of accuracy, the implementation variant of the cascade ensemble shown in Figure 3 improves as the three randomly formed subsamples with repetitions grow based on the training dataset. However, this trend is observed only until they reach a size of 70% of the training set.
- Using subsamples sized at 80% or more leads to a significant reduction in error in the application mode, likely due to the limited generalization properties of such a cascade ensemble variant.
- The highest accuracy in the application mode in Figure 2(b) and the most robust generalization properties in Figure 3 of the enhanced cascade ensemble were attained when employing a three-level cascade structure, with each of the three subsamples comprising 70% randomly chosen vectors with repetitions from the training dataset. The disparity in accuracy between training and application modes is minimal in this scenario. Thus, these parameters represent the optimal operation configuration for the enhanced cascade ensemble on the given dataset.

Based on the selected optimal parameters of the enhanced cascade ensemble, performance indicators of its operation in both modes are compiled in Table 1. The F1-score values from Table 1 were used to compare the effectiveness of the enhanced cascade ensemble with a variety of existing methods.

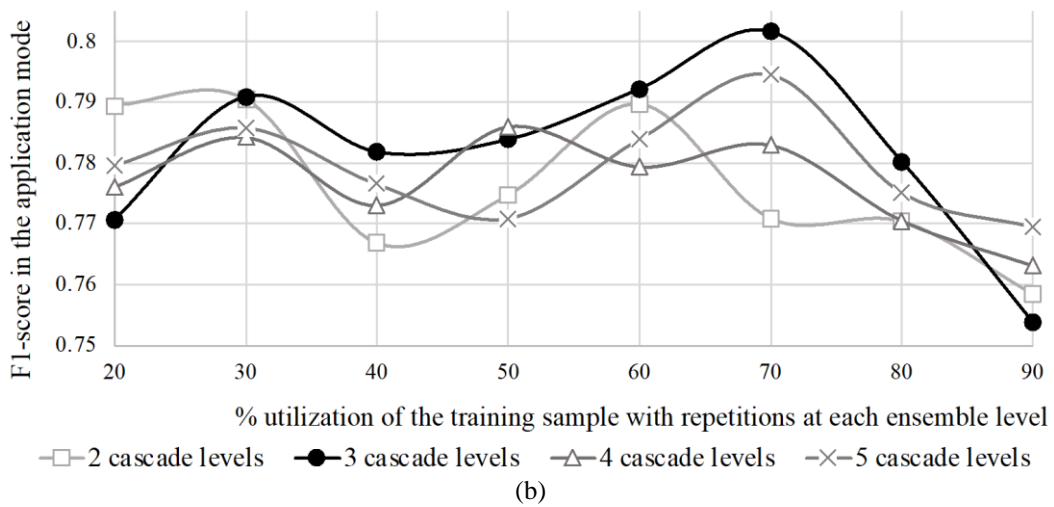
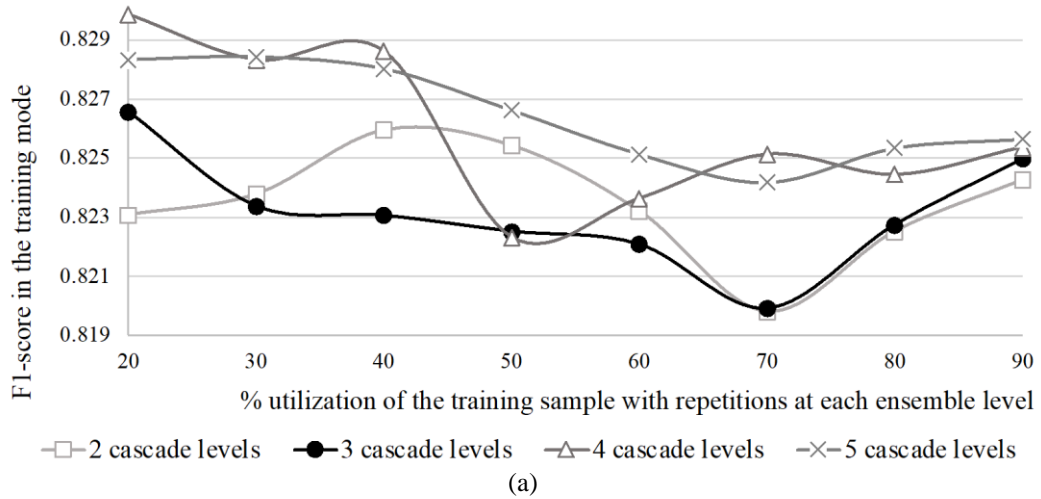


Figure 2. The investigation of the enhanced cascade ensemble accuracy while varying both the percentage of training sample utilization with repetition and the depth of the cascade for (a) the training mode and (b) the application mode

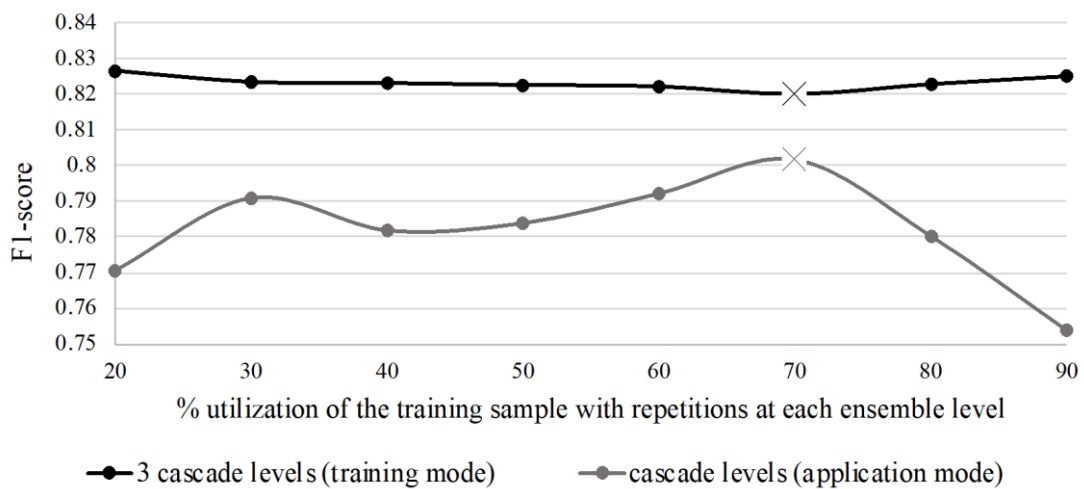


Figure 3. The results of the enhanced cascade ensemble with three levels when changing the percentage of using training samples for each of the three subsamples

Table 1. Results of the enhanced cascade ensemble

Performance indicator	Training mode	Test mode
Precision	0.831	0.879
Recall	0.821	0.751
F1-score	0.819	0.802
Training time (seconds)	3.35	-

### 3.3. Comparison and discussion

The effectiveness comparison of the enhanced cascade ensemble was conducted with a series of such methods:

- SGD algorithm [39], as the method underlying the operation of the cascade ensemble;
- The method from [40], as a method from [39] which, like the cascade ensemble, utilizes Ito decomposition to improve the accuracy of SGD operation;
- Initial cascade ensemble [33], as the baseline method improved upon in this article.
- The comparison results based on accuracy, generalization properties, and training procedure duration are summarized in Figure 4, with Figure 4(a) emphasizing the comparison based on the F1-score.

From Figure 4, it can be observed that the method [39] demonstrates the lowest accuracy. However, it shows the shortest training time in Figure 4(b). The method from [40] exhibits significantly higher accuracy, albeit slightly poorer generalization. This can be attributed to the use of Ito decomposition for nonlinear expansion of inputs. Such an approach has led to a substantial increase in accuracy, explained by the Cover's theorem. However, due to the replacement of the initial 29 inputs of the problem with 435 inputs obtained after using Ito decomposition, this method has decreased its generalization properties and, worst of all, increased the training procedure duration by almost 36 times.

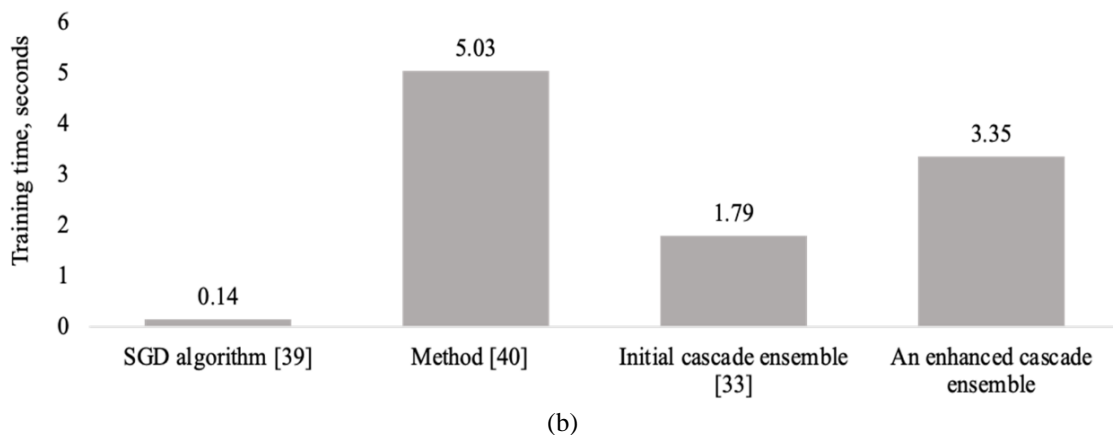
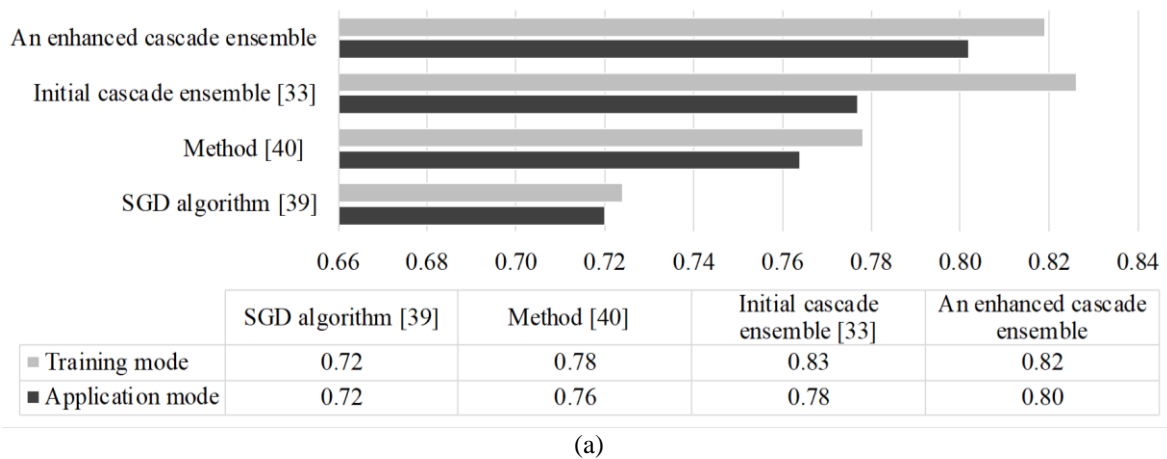


Figure 4. Comparison of the results for all methods based on (a) the F1-score and (b) the duration of the training procedure (in seconds)



The baseline cascade ensemble [33] demonstrated higher accuracy compared to both previous methods. However, due to the significant increase in independent attributes of the problem caused by the use of second-degree Ito decomposition, it exhibits the poorest generalization properties compared to all investigated methods. Nevertheless, distributing the large dataset into unique subsets for each classifier of this cascade, significantly reduced the training time of the method.

The enhanced cascade ensemble showed the best results in terms of accuracy. Additionally, the new procedure proposed in this study for forming subsets for each cascade level resulted in a significant improvement in the generalization properties of the enhanced method. Specifically, the difference between the training and application accuracy of this method is only 2%, whereas in the baseline method, it is almost 5%. It provides advantages for using this method of practice.

However, using larger subsets at each level of the cascade ensemble increased its training time (almost two times). Therefore, in the prospects of further research, attention should be focused on reducing the training time of this method. This can be achieved, in particular, by using various dimensionality reduction methods. Such an approach seems most promising due to the necessity of using Ito decomposition in the cascade, which, on the one hand, significantly increases the accuracy of linear classifiers but, on the other hand, greatly increases the dimensionality of the problem. The approach using, for example, principal component analysis (PCA) should eliminate this drawback. Additionally, in the future, the possibility of using more accurate linear methods as weak predictors of the cascade, particularly from the class of neural networks, should be considered.

#### 4. CONCLUSION

This paper underscores the indispensable role of biomedical data classification in modern medicine amidst rapid technological and healthcare advancements. While single models with optimal parameters may fall short in achieving satisfactory classification accuracy, ensembles of machine learning or artificial neural networks methods offer a promising solution. Cascaded ensembles, in particular, are highlighted as one of the most accurate classes of ensemble methods, characterized by high accuracy and strong generalization properties. These models employ hierarchical training, where each model utilizes the outputs of preceding models in the hierarchy. However, existing cascade classifiers often suffer from limitations, such as incomplete information at each level due to the equal division of large input datasets. To address this issue, we propose an enhanced cascaded ensemble scheme using a novel data sampling partitioning approach. This approach forms significantly larger subsamples at each cascade level, enabling artificial intelligent means to access more data for training procedures, which is particularly beneficial for analyzing large and medium-volume biomedical data. Modeling on real-world data and experimental comparisons between existing and modified cascade ensembles demonstrate a substantial increase in classification accuracy and improved generalization properties with our proposed approach. Moving forward, future research will focus on developing methods to reduce the duration of the cascade training procedure while maintaining accuracy, considering the use of the Ito decomposition at each level, which significantly increases the dimensionality of the input data space.

#### ACKNOWLEDGEMENTS

This work was funded by the National Research Foundation of Ukraine, project number 30/0103. The authors thank the British Academy's Researchers at Risk Fellowships Program which supports this paper.

#### REFERENCES




- [1] A. Altameem, V. Kovtun, M. Al-Ma'aitah, T. Altameem, F. H., and A. E. Youssef, "Patient's data privacy protection in medical healthcare transmission services using back propagation learning," *Computers and Electrical Engineering*, vol. 102, Sep. 2022, doi: 10.1016/j.compeleceng.2022.108087.
- [2] D. Srivastava, H. Pandey, and A. K. Agarwal, "Complex predictive analysis for health care: a comprehensive review," *Bulletin of Electrical Engineering and Informatics*, vol. 12, no. 1, pp. 521–531, Feb. 2023, doi: 10.11591/eei.v12i1.4373.
- [3] B. T. Ahmed, "Data mining techniques for lung and breast cancer diagnosis: A review," *International Journal of Informatics and Communication Technology*, vol. 10, no. 2, Aug. 2021, doi: 10.11591/ijict.v10i2.pp93-103.
- [4] U. K. Hisan and M. M. Amri, "Recommendation of precision medicine application in Indonesia from multiple perspective: a review," *International Journal of Public Health Science*, vol. 12, no. 1, pp. 225–238, Mar. 2023, doi: 10.11591/ijphs.v12i1.22010.
- [5] Y. Tolstyak, V. Chopyak, and M. Havryliuk, "An investigation of the primary immunosuppressive therapy's influence on kidney transplant survival at one month after transplantation," *Transplant Immunology*, vol. 78, Jun. 2023, doi: 10.1016/j.trim.2023.101832.
- [6] Y. Tolstyak and M. Havryliuk, "An assessment of the transplant's survival level for recipients after kidney transplantations using cox proportional-hazards model," *IDDM-2022: 5th International Conference on Informatics & Data-Driven Medicine*, vol. 3302, 2022.

- [7] C. Jittawiriyankoon, "Performance evaluation of listwise deletion for impaired datasets in multiple regression-based prediction," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 15, no. 2, pp. 1009–1018, Aug. 2019, doi: 10.11591/ijeecs.v15.i2.pp1009-1018.
- [8] A. G. Shaaban, M. H. Khafagy, M. A. Elmasry, H. El-Beih, and M. H. Ibrahim, "Knowledge discovery in manufacturing datasets using data mining techniques to improve business performance," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 26, no. 3, pp. 1736–1746, Jun. 2022, doi: 10.11591/ijeecs.v26.i3.pp1736-1746.
- [9] O. A. Alade, R. Sallehuddin, and N. H. M. Radzi, "Enhancing extreme learning machines classification with moth-flame optimization technique," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 26, no. 2, pp. 1027–1035, May 2022, doi: 10.11591/ijeecs.v26.i2.pp1027-1035.
- [10] D. Chumachenko, P. Piletskiy, M. Sukhorukova, and T. Chumachenko, "Predictive model of Lyme disease epidemic process using machine learning approach," *Applied Sciences*, vol. 12, no. 9, Apr. 2022, doi: 10.3390/app12094282.
- [11] D. Chumachenko, O. Sokolov, and S. Yakovlev, "Fuzzy recurrent mappings in multiagent simulation of population dynamics systems," *International Journal of Computing*, pp. 290–297, Jun. 2020, doi: 10.47839/ijc.19.2.1773.
- [12] N. Krishnados and L. Kumar Ramasamy, "A study on high dimensional big data using predictive data analytics model," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 30, no. 1, pp. 174–182, Apr. 2023, doi: 10.11591/ijeecs.v30.i1.pp174-182.
- [13] A. Abdullahi, M. Ali Barre, and A. Hussein Elmi, "A machine learning approach to cardiovascular disease prediction with advanced feature selection," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 33, no. 2, pp. 1030–1041, Feb. 2024, doi: 10.11591/ijeecs.v33.i2.pp1030-1041.
- [14] N. R. Kolukula, P. N. Pothineni, V. M. K. Chinta, V. G. Boppana, R. P. Kalapala, and S. Duvvi, "Predictive analytics of heart disease presence with feature importance based on machine learning algorithms," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 32, no. 2, pp. 1070–1077, Nov. 2023, doi: 10.11591/ijeecs.v32.i2.pp1070-1077.
- [15] J. Bilski, J. Smoląg, B. Kowalczyk, K. Grzanek, and I. Izonin, "Fast computational approach to the Levenberg-Marquardt algorithm for training feedforward neural networks," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 13, no. 2, pp. 45–61, Mar. 2023, doi: 10.2478/jaiscr-2023-0006.
- [16] L. Mochurad, Y. Hladun, Y. Zasoba, and M. Gregus, "An approach for opening doors with a mobile robot using machine learning methods," *Big Data and Cognitive Computing*, vol. 7, no. 2, Apr. 2023, doi: 10.3390/bdcc7020069.
- [17] S. Chopparapu and B. Seventline Joseph, "A hybrid facial features extraction-based classification framework for typhlotic people," *Bulletin of Electrical Engineering and Informatics*, vol. 13, no. 1, pp. 338–349, Feb. 2024, doi: 10.11591/eei.v13i1.5628.
- [18] I. Perova and Y. Bodyanskiy, "Adaptive human machine interaction approach for feature selection-extraction task in medical data mining," *International Journal of Computing*, pp. 113–119, Jun. 2018, doi: 10.47839/ijc.17.2.997.
- [19] N. Santoso, W. Wibowo, and H. Hikmawati, "Integration of synthetic minority oversampling technique for imbalanced class," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 13, no. 1, pp. 102–108, Jan. 2019, doi: 10.11591/ijeecs.v13.i1.pp102-108.
- [20] A. J. Barid, H. Hadiyanto, and A. Wibowo, "Optimization of the algorithms use ensemble and synthetic minority oversampling technique for air quality classification," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 33, no. 3, pp. 1632–1640, Mar. 2024, doi: 10.11591/ijeecs.v33.i3.pp1632-1640.
- [21] L. Zomchak and V. Melnychuk, "Creditworthiness of individual borrowers forecasting with machine learning methods," in *International Conference of Artificial Intelligence, Medical Engineering, Education*, 2023, pp. 553–561.
- [22] E. Elvin and A. Wibowo, "Forecasting water quality through machine learning and hyperparameter optimization," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 33, no. 1, pp. 496–506, Jan. 2024, doi: 10.11591/ijeecs.v33.i1.pp496-506.
- [23] C. Ganguli, S. K. Shandilya, M. Nehrey, and M. Havryliuk, "Adaptive artificial bee colony algorithm for nature-inspired cyber defense," *Systems*, vol. 11, no. 1, Jan. 2023, doi: 10.3390/systems11010027.
- [24] A. Adeleke, N. A. Samsudin, M. H. Abdul Rahim, S. K. Ahmad Khalid, and R. Efendi, "Multi-label classification approach for quranic verses labeling," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 24, no. 1, pp. 484–490, Oct. 2021, doi: 10.11591/ijeecs.v24.i1.pp484-490.
- [25] V. Yakovyna and N. Shakhovska, "Software failure time series prediction with RBF, GRNN, and LSTM neural networks," *Procedia Computer Science*, vol. 207, pp. 837–847, 2022, doi: 10.1016/j.procs.2022.09.139.
- [26] D. Pankaj Javale and S. Suhas Desai, "Machine learning ensemble approach for healthcare data analytics," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 28, no. 2, pp. 926–933, Nov. 2022, doi: 10.11591/ijeecs.v28.i2.pp926-933.
- [27] L. Mochurad and G. Shchur, "Parallelization of cryptographic algorithm based on different parallel computing technologies," *IT&AS'2021: Symposium on Information Technologies & Applied Sciences*, pp. 20–29, 2021.
- [28] V. D. Babu and K. Malathi, "Large dataset partitioning using ensemble partition-based clustering with majority voting technique," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 29, no. 2, pp. 838–844, Feb. 2023, doi: 10.11591/ijeecs.v29.i2.pp838-844.
- [29] O. Mulesa, V. Snytyuk, and I. Myronyuk, "Optimal alternative selection models in a multi-stage decision-making process," *EUREKA: Physics and Engineering*, vol. 6, pp. 43–50, Nov. 2019, doi: 10.21303/2461-4262.2019.001005.
- [30] O. Mulesa, F. Geche, A. Batyuk, and V. Buchok, "Development of combined information technology for time series prediction," in *Conference on Computer Science and Information Technologies*, 2018, pp. 361–373.
- [31] I. de Zarzà, J. de Curtò, E. Hernández-Orallo, and C. T. Calafate, "Cascading and ensemble techniques in deep learning," *Electronics*, vol. 12, no. 15, Aug. 2023, doi: 10.3390/electronics12153354.
- [32] S. Xu, Q. Tang, L. Jin, and Z. Pan, "A cascade ensemble learning model for human activity recognition with smartphones," *Sensors*, vol. 19, no. 10, May 2019, doi: 10.3390/s19102307.
- [33] I. Izonin, R. Tkachenko, R. Holoven, K. Yemets, M. Havryliuk, and S. K. Shandilya, "SGD-based cascade scheme for higher degrees wiener polynomial approximation of large biomedical datasets," *Machine Learning and Knowledge Extraction*, vol. 4, no. 4, pp. 1088–1106, Nov. 2022, doi: 10.3390/make4040055.
- [34] G. V. Gopal and G. R. M. Babu, "An ensemble feature selection approach using hybrid kernel based SVM for network intrusion detection system," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 23, no. 1, pp. 558–565, Jul. 2021, doi: 10.11591/ijeecs.v23.i1.pp558-565.
- [35] I. Izonin, R. Tkachenko, O. Gurbych, M. Kovac, L. Rutkowski, and R. Holoven, "A non-linear SVR-based cascade model for improving prediction accuracy of biomedical data analysis," *Mathematical Biosciences and Engineering*, vol. 20, no. 7, pp. 13398–13414, 2023, doi: 10.3934/mbe.2023597.




- [36] A. A. Khan, O. Chaudhari, and R. Chandra, "A review of ensemble learning and data augmentation models for class imbalanced problems: combination, implementation and evaluation," *Expert Systems with Applications*, vol. 244, Jun. 2024, doi: 10.1016/j.eswa.2023.122778.
- [37] "CDC - 2021 BRFSS survey data and documentation," [www.cdc.gov](https://www.cdc.gov/brfss/annual_data/annual_2021.html), [https://www.cdc.gov/brfss/annual\\_data/annual\\_2021.html](https://www.cdc.gov/brfss/annual_data/annual_2021.html) (accessed Mar. 04, 2024).
- [38] A. Sambir, V. Yakovyna, and M. Seniv, "Recruiting software architecture using user generated data," in *2017 XIIIth International Conference on Perspective Technologies and Methods in MEMS Design (MEMSTECH)*, Apr. 2017, pp. 161–163, doi: 10.1109/MEMSTECH.2017.7937557.
- [39] "SGDRegressor," [scikit-learn.org](https://scikit-learn.org), [https://scikit-learn.org/stable/modules/generated/sklearn.linear\\_model.SGDRegressor.html](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.SGDRegressor.html) (accessed Feb. 08, 2019).
- [40] I. Izonin, M. Greguš ml., R. Tkachenko, M. Logoyda, O. Mishchuk, and Y. Kynash, "SGD-based wiener polynomial approximation for missing data recovery in air pollution monitoring dataset," in *International Work-Conference on Artificial Neural Networks*, 2019, pp. 781–793.

## BIOGRAPHIES OF AUTHORS






**Ivan Izonin**    is an associate professor at the Department of Artificial Intelligence of Lviv Polytechnic National University, Ukraine, and also a research fellow at the Department of Civil Engineering, School of Engineering, University of Birmingham, UK. He received an MSc degree in computer science from Lviv Polytechnic National University, Ukraine in 2011, and an MSc degree in economic cybernetics from the Ivan Franko National University of Lviv, Ukraine, in 2012. He received a Ph.D. degree in artificial intelligence from the Institute of Computer Science and Information Technologies of Lviv Polytechnic National University, Ukraine in 2016. His research interests include small biomedical data analysis, meta-learning, and ensemble methods using no-iterative training algorithms where he is the author/co-author of over 150 research publications. He can be contacted at email: [ivan.v.izonin@gmail.com](mailto:ivan.v.izonin@gmail.com) or [i.izonin@bham.ac.uk](mailto:i.izonin@bham.ac.uk).






**Roman Muzyka**    is a Ph.D. student at the Department of Artificial Intelligence of Lviv Polytechnic National University, Ukraine, and works as a data engineer. He received his master of science (M.Sc.) degree in computer science from Lviv Polytechnic National University, Ukraine in 2023. His primary research interests lie in ensemble methods, cascades, machine learning, and data analysis. Additionally, Roman was a member of the Junior Academy of Sciences of Ukraine from 2016 to 2018, actively contributing to various scientific initiatives. He also participated in the international Erasmus+ projects, where he focused on robotics, programming, and mathematical analysis. For inquiries, he can be reached at [roman.muzyka.mknssh.2022@lpnu.ua](mailto:roman.muzyka.mknssh.2022@lpnu.ua).







**Roman Tkachenko**    is a professor at the Department of Publishing Information Technologies of Lviv Polytechnic National University, Ukraine since 2002. From 2017-2020, he was also the Head of this department. He received an MSc in electrical engineering from Lviv Polytechnic National University, Ukraine in 1972, and a Ph.D. in metrology from the Lviv Polytechnic National University, Ukraine in 1984. Prof. Tkachenko received his Doctor of Science degree in information technologies from the State Research Institute of Information Infrastructure, Lviv, Ukraine. His research interests are primarily in computational intelligence, high-speed neural-like systems, non-iterative machine learning algorithms, and ensemble learning. He is the author/co-author of over 200 research publications. He can be contacted at [roman.tkachenk@gmail.com](mailto:roman.tkachenk@gmail.com).







**Michal Gregus**    is professor of the Faculty of Management, Comenius University Bratislava, Bratislava, Slovak Republic. He finished his university studies with summa cum laude and obtained his PhD degree in the field of mathematical analysis at the Faculty of Mathematics and Physics at Comenius University in Bratislava. He has been working previously in the field of functional analysis and its applications. At present, his research interests are in management information systems, in modelling of economic processes and in business analytics. He can be contacted at [Michal.Gregus@fm.uniba.sk](mailto:Michal.Gregus@fm.uniba.sk).



**Natalya Kustra**     is an associate professor of the Department of Information Technologies of Publishing at the Lviv Polytechnic National University, Ukraine. She received her master's degree in computer science from Lviv Polytechnic National University, Ukraine, in 2002. She received a Ph.D. in artificial intelligence at the Institute of Computer Sciences and Information Technologies of Lviv Polytechnic National University, Ukraine in 2008. Her research interests include image dataset analysis, meta-learning, and ensemble methods using iteration-free learning algorithms, where she has authored/co-authored more than 30 scientific publications. She can be contacted at email: natalia.o.kustra@lpnu.ua.



**Stergios-Aristoteles Mitoulis**     is the leader of MetaInfrastructure.org and www.bridgeUkraine.org. He is the head of structures at the University of Birmingham and associate professor with a focus on infrastructure safety, resilience, and digitalization. Delivering integrated solutions to climate hazards and multiple stressors. He holds a 5-year Diploma from Aristotle University and MSc in sustainable design of infrastructure to natural hazards and a PhD in 2008. He is CEng MICE, a member of the ASCE, EAEE, IABSE and FHEA-UK. He has published more than 130 journal and conference papers. Stergios has worked as PI, Co-I and researcher for more than 25 research projects, including coordinating and leading a 1.7m-Euro HORIZON-MSCA-SE multi-partner project on climate-aware resilience for sustainable and interdependent infrastructure systems. He can be contacted at email: s.a.mitoulis@bham.ac.uk.