REVIEW ARTICLE

CONTENTS

# Artificial intelligence in disease diagnosis: Application within feature and state spaces

**Ozar P. Mintser, Larysa Yu. Babintseva, Stanislav I. Mokhnachov, Olga O. Sukhanova**
SHUPYK NATIONAL UNIVERSITY OF HEALTHCARE OF UKRAINE, KYIV, UKRAINE

## ABSTRACT

**Aim:** To justify strategies for transitioning from the topology of state and feature spaces to a unified, structured metric space within medical decision-making.

**Materials and Methods:** The paper proposes the simultaneous use of two conceptual spaces-states and features - when applying artificial intelligence in healthcare to support evidence-based diagnostic decision-making. To implement classification mechanisms using artificial intelligence, it is proposed to use the Multiscale Classifier algorithm with additional spatial analysis to extract features and evaluate the quality of the obtained classifications using ROC curves (Receiver Operating Characteristic Curves). For modeling the state space, it is effective to use recursive Bayesian state estimation, dynamic regression, and correlation analysis.

**Conclusions:** Findings indicate that the accuracy and efficiency of classification processes can be substantially enhanced by adopting dynamic classification principles. Moreover, the overall effectiveness of AI deployment in healthcare partially depends on the extent of generalization and the specific structural organization of medical information. A critical and practical element at the pre-implementation stage of AI integration is the development of a domain-specific ontology.

**KEY WORDS:** artificial intelligence, diagnosis and prognosis of diseases, optimization algorithms, naive algorithms, hybrid approaches

Wiad Lek. 2025;78(7):1411-1417. doi: 10.36740/WLek/208997 DOI

## INTRODUCTION

In recent years, there has been remarkable progress in artificial intelligence (AI) research and its application in practical medicine, leading to a heightened interest in decision-making challenges. Many researchers report significant success in recognizing pathological processes [1-4]. This success is attributed mainly to advances in neural networks and profound learning principles. A key feature of these networks is their ability to self-learn, which represents an essential aspect of the AI state space [5]. Another vital aspect is generalization, defined as the ability to apply knowledge from specific cases to more general problems. However, several serious challenges remain. One of the primary issues is the difficulty of providing clear explanations for diagnoses, which can be unsatisfying for medical professionals. Applying clinical precedent logic does not always meet the needs of specialists. In previous work, we proposed using the "Ex juvantibus" strategy, which theoretically could lead to more effective outcomes [6]. Nonetheless, this approach is not universally applicable, mainly because it can be challenging to predict the dynamics of a pathological process.

Significant difficulties are also associated with the principles of semantic justification of the conclusion due to information correlation, lack of uniform classifications, etc.

## AIM

The study aims to substantiate the transition strategy from the topology of the feature and state space to a single structured metric space for decision-making.

## MATERIALS AND METHODS

The evolution of computer-assisted disease diagnosis spans decades, beginning with the foundational works of Ledley and Lusted [7, 8] in the mid-20th century. Since then, substantial progress has been made, with the modern era marked by the successful application of machine learning (ML) methods as a central driver of diagnostic innovation. Currently, a broad array of automatic classification techniques is available, supported by foundational ML methodologies such as clinical precedent-based approaches, hyperplane, and hyper-

rectangle strategies, fuzzy logic systems, probabilistic algorithms (e.g., the Bayesian framework and Wald's sequential statistical analysis), as well as perceptron and multilayer perceptron models.

This study introduces a novel hyperrectangle-based algorithm referred to as the Multiscale Classifier (MSC), implemented through an inductive decision tree framework [9]. The MSC algorithm applies to N-dimensional classification problems by recursively dividing the feature space in half, with tree growth regulated via logical minimization. Subsequent stages in the classification process necessitate decision trees that account for the cost of misclassification. It has been demonstrated that such models support diverse classification operating modes, which can be visualized through ROC curves. The MSC offers several notable advantages over conventional hyperrectangle-based methods: it enables incremental learning, utilizes a non-binary tree structure, and allows reverse decision propagation, among other enhancements.

Additionally, a feature extraction methodology based on scale-spatial analysis was proposed and applied to selected diagnostic indicators. Empirical findings suggest that this approach yields improved performance compared to traditional feature extraction techniques.

A complementary investigation evaluated the performance of various machine learning algorithms across six "real-world" medical diagnostic datasets [10]. Each algorithm was assessed following principles of evidence-based medicine, using metrics such as overall accuracy, sensitivity, specificity, area under the ROC curve (AUC), chi-square test statistics, training time, and interpretability. Analysis of variance (ANOVA) was employed to assess the statistical significance of observed differences in cross-validated accuracy and AUC outcomes.

The findings highlight the advantages of AUC over precision, particularly its higher statistical sensitivity, threshold independence, and robustness to variations in prior class probabilities. While exemplar-based and hyperplane-oriented methods showed slightly higher classification accuracy, hyperrectangle-based approaches—including MSC—offered superior interpretability and required fewer computational resources. The Multiscale Classifier demonstrated competitive performance among evaluated models. The authors propose that MSC holds strong potential as a supplementary tool for enhancing diagnostic accuracy in medical machine-learning applications.

It is also essential to briefly discuss the role of so-called naive algorithms. These approaches are typically defined by their simplicity and reliance on suboptimal or heuristic-based problem-solving strategies. Often, they employ straightforward procedures without leveraging advanced optimization techniques or computational refinements.

Study [1] underscores that standard datasets related to chronic diseases can be compiled from various global sources. However, datasets about specific chronic conditions frequently contain ambiguous class instances—cases that exhibit features representative of two or more diagnostic categories. Such ambiguity introduces classification challenges and increases the likelihood of reduced model performance in machine learning systems. A key contribution of the referenced study is the incorporation of fuzzy clustering, which is implemented through the method of rough averages. The study evaluated several machine learning algorithms, including the naive Bayes (NB) classifier, Boltzmann machine, k-nearest neighbor (kNN), support vector machine (SVM), decision tree, and logistic regression models. The dataset used for analysis was drawn from a widely recognized machine-learning repository focused on chronic disease data. Experimental results confirmed that the proposed system effectively supports chronic disease diagnostics. Among the evaluated algorithms, the naive Bayes classifier achieved the highest performance in classifying diabetic conditions, with an accuracy rate of 80.55%. For relatively straightforward clinical scenarios, naive algorithms deliver reliable outcomes and, due to their computational efficiency and low implementation cost, may be considered optimal in such contexts.

Nonetheless, regardless of the mathematical strengths of certain classification approaches, significant challenges persist regarding the selection of relevant features associated with pathological processes. These challenges stem largely from the substantial variability in clinical presentations, the coexistence of multiple disease combinations, and the resultant overlap in symptomatology. For instance, the combinatorial possibilities of clinical features across just 100 pathological processes yield millions of potential disease scenarios. Consequently, assembling a sufficiently robust statistical foundation for algorithm training becomes a formidable task, even when employing globally accessible, open healthcare data systems.

## REVIEW AND DISCUSSION

The evolution of computer-assisted disease diagnosis spans decades, beginning with the foundational works of Ledley and Lusted [7, 8] in the mid-20th century. Since then, substantial progress has been made, with the modern era marked by the successful application of machine learning (ML) methods as a central driver of

diagnostic innovation. Currently, a broad array of automatic classification techniques is available, supported by foundational ML methodologies such as clinical precedent-based approaches, hyperplane, and hyperrectangle strategies, fuzzy logic systems, probabilistic algorithms (e.g., the Bayesian framework and Wald's sequential statistical analysis), as well as perceptron and multilayer perceptron models.

This study introduces a novel hyperrectangle-based algorithm referred to as the Multiscale Classifier (MSC), implemented through an inductive decision tree framework [9]. The MSC algorithm applies to N-dimensional classification problems by recursively dividing the feature space in half, with tree growth regulated via logical minimization. Subsequent stages in the classification process necessitate decision trees that account for the cost of misclassification. It has been demonstrated that such models support diverse classification operating modes, which can be visualized through Receiver Operating Characteristic (ROC) curves. The MSC offers several notable advantages over conventional hyperrectangle-based methods: it enables incremental learning, utilizes a non-binary tree structure, and allows reverse decision propagation, among other enhancements.

Additionally, a feature extraction methodology based on scale-spatial analysis was proposed and applied to selected diagnostic indicators. Empirical findings suggest that this approach yields improved performance compared to traditional feature extraction techniques.

A complementary investigation evaluated the performance of various machine learning algorithms across six "real-world" medical diagnostic datasets [10]. Each algorithm was assessed following principles of evidence-based medicine, using metrics such as overall accuracy, sensitivity, specificity, area under the ROC curve (AUC), chi-square test statistics, training time, and interpretability. Analysis of variance (ANOVA) was employed to assess the statistical significance of observed differences in cross-validated accuracy and AUC outcomes.

The findings highlight the advantages of AUC over precision, particularly its higher statistical sensitivity, threshold independence, and robustness to variations in prior class probabilities. While exemplar-based and hyperplane-oriented methods showed slightly higher classification accuracy, hyperrectangle-based approaches—including MSC—offered superior interpretability and required fewer computational resources. The Multiscale Classifier demonstrated competitive performance among evaluated models. The authors propose that MSC holds strong potential as a supplementary tool for enhancing diagnostic accuracy in medical machine-learning applications.

It is also essential to briefly discuss the role of so-called naive algorithms. These approaches are typically defined by their simplicity and reliance on suboptimal or heuristic-based problem-solving strategies. Often, they employ straightforward procedures without leveraging advanced optimization techniques or computational refinements.

Study [1] underscores that standard datasets related to chronic diseases can be compiled from various global sources. However, datasets about specific chronic conditions frequently contain ambiguous class instances—cases that exhibit features representative of two or more diagnostic categories. Such ambiguity introduces classification challenges and increases the likelihood of reduced model performance in machine learning systems. A key contribution of the referenced study is the incorporation of fuzzy clustering, which is implemented through the method of rough averages. The study evaluated several machine learning algorithms, including the naive Bayes (NB) classifier, Boltzmann machine, k-nearest neighbor (kNN), support vector machine (SVM), decision tree, and logistic regression models. The dataset used for analysis was drawn from a widely recognized machine-learning repository focused on chronic disease data. Experimental results confirmed that the proposed system effectively supports chronic disease diagnostics. Among the evaluated algorithms, the naive Bayes classifier achieved the highest performance in classifying diabetic conditions, with an accuracy rate of 80.55%. For relatively straightforward clinical scenarios, naive algorithms deliver reliable outcomes and, due to their computational efficiency and low implementation cost, may be considered optimal in such contexts.

Nonetheless, regardless of the mathematical strengths of certain classification approaches, significant challenges persist regarding the selection of relevant features associated with pathological processes. These challenges stem largely from the substantial variability in clinical presentations, the coexistence of multiple disease combinations, and the resultant overlap in symptomatology. For instance, the combinatorial possibilities of clinical features across just 100 pathological processes yield millions of potential disease scenarios. Consequently, assembling a sufficiently robust statistical foundation for algorithm training becomes a formidable task, even when employing globally accessible, open healthcare data systems.

## USING OF THE FEATURE SPACE

Pattern recognition in biomedical datasets often proves challenging due to the presence of numerous irrelevant

or redundant attributes. Implementing a robust feature selection (FS) strategy is essential to eliminate non-informative or redundant variables. The primary objective of FS methods is to enhance classification performance by reducing dimensionality and retaining only the most relevant features [4]. Accurate execution of the data cleaning phase is pivotal to achieving optimal machine learning outcomes. This empirical investigation presents a classification methodology for biomedical data using FS techniques. The proposed framework incorporates three soft computing-based optimization algorithms: Teaching–Learning-Based Optimization (TLBO), Elephant Herding Optimization (EHO), and a hybridized approach that combines both strengths.

EHO is a relatively novel swarm-based metaheuristic search algorithm inspired by the social behavior of elephant herds [11]. One of the first applications of this method was to solve general optimization problems. The method reflects two different behavioral stages in the social dynamics of elephants: in nature, elephants belonging to different clans live together, but when they become adults, they leave their family group. These phenomena are modeled using two operators - the group update operator and the separation operator - both of which increase the diversity of the population at later stages of the search. The cited study [11] demonstrated that EHO performs efficiently when compared with other established metaheuristic algorithms.

Although the aforementioned algorithms have been employed in previous research, their efficacy in solving FS challenges—particularly in disease prognosis—has not only been validated but also extended into new application domains. For example, the referenced study [4] assessed classification performance in differentiating benign and malignant tumors using the publicly available Wisconsin Diagnostic Breast Cancer (WDBC) dataset. To mitigate the risk of overfitting, five-fold cross-validation was employed. Evaluation metrics included sensitivity, specificity, accuracy, precision, reliability, and area under the receiver operating characteristic curve (AUC). The highest recorded classification accuracy using the proposed FS approach was 97.96%.

USING THE STATE SPACE

Tasks framed within the state space differ markedly from those in the feature space. Recent research has focused on managing fuzzy boundaries between states [12]. For instance, a breast cancer diagnostic algorithm has been proposed that consists of a two-part structure. The first part includes four sequential steps: preprocessing of image data; image analysis using wavelet transformation; feature extraction using wavelet-derived parameters to isolate the most significant characteristics via standard separation methods; classification using fuzzy logic to determine whether the tumor is benign or malignant. This algorithm has demonstrated a classification accuracy of 98%.

A persistent challenge in this domain is the dynamic nature of disease characteristics over time. Consequently, state-space modeling offers a compelling framework for estimating the evolving states of dynamic biological systems. The application of recursive Bayesian state estimation enables efficient model updates based on new observations, making this approach particularly well-suited to settings characterized by noisy or incomplete data. Christopher D. Prashad employed this methodology in his research on infectious disease modeling [13]. The primary objective was to enhance model accuracy and computational efficiency when applying state space modeling principles to epidemic data. The study addressed key issues such as parameter uncertainty and identification logic, which are critical in the context of epidemiological forecasting [14]. Compared to traditional frequency-based models, Bayesian approaches offer a more robust mechanism for incorporating uncertainty into predictive models, thereby deepening our understanding of disease dynamics.

The study further demonstrated the utility of state space modeling, recursive Bayesian estimation, dynamic regression, and correlation analysis for analyzing public health data. These tools significantly enhance both descriptive and predictive capabilities when modeling dynamic systems. The study results demonstrate the ability of the models to recognize key epidemiological trends.

The authors underline that future work could explore ways to optimize the algorithm for applications with real-time or streaming data, focusing on reducing computational complexity. In addition, practical implementation and algorithmic integration into real-time systems such as patient monitoring, long-term management of pathological conditions, and clinical decision support are promising areas for further development.

The recursive Bayesian approach to state estimation represents a significant advancement in state space modeling. Its inherent adaptability to changing conditions, coupled with its capacity to maintain high computational accuracy, renders it particularly valuable in scenarios requiring precise state estimation under conditions of uncertainty. Nevertheless, several challenges remain. These include the temporal variability of clinical manifestations of diseases, errors in identifying key pathological indicators, and the difficulty of quickly comparing the feature space with the state space. To solve these problems, we are currently considering the integration of ontological tools into the technological process.

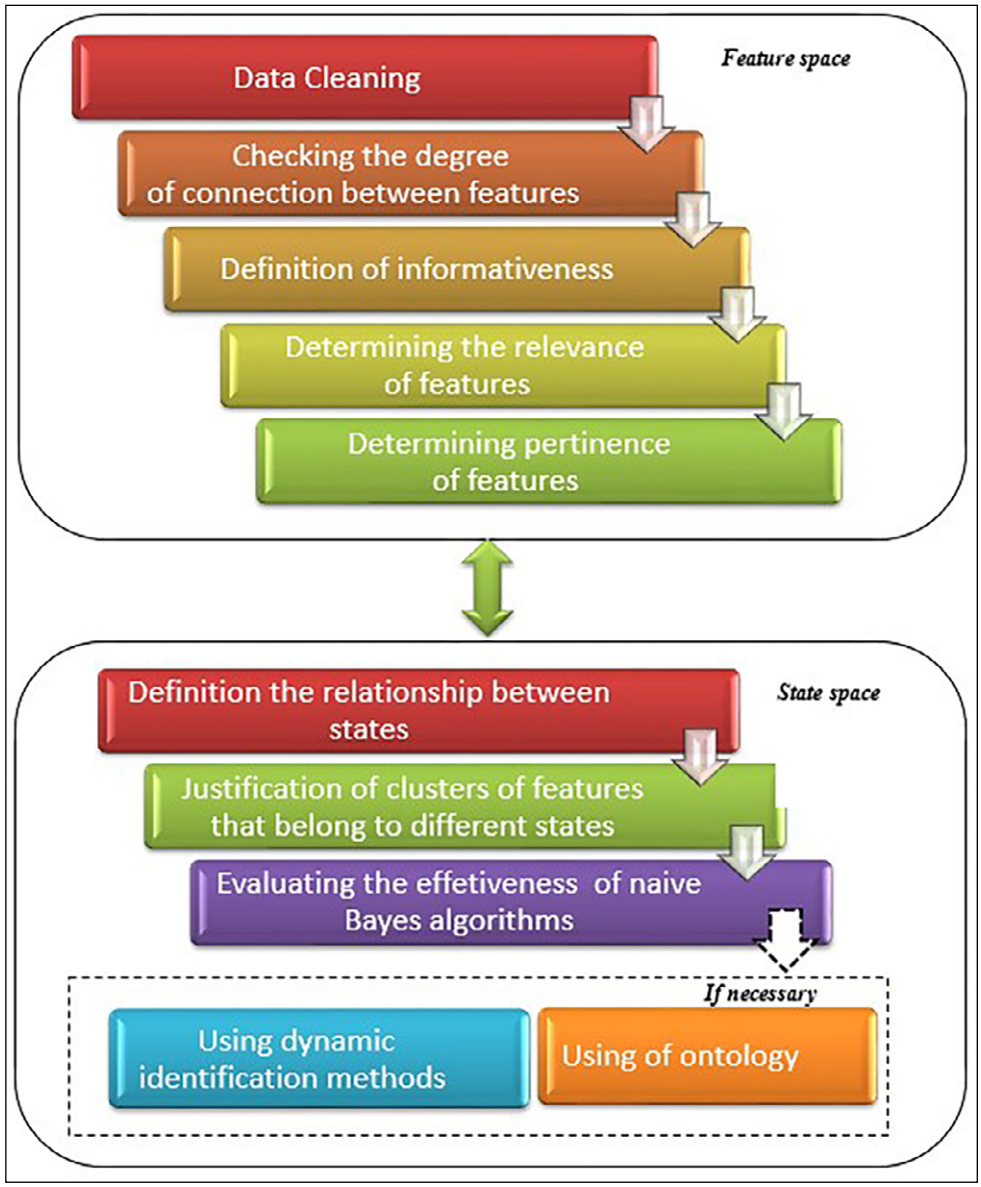


**Fig. 1.** General Scheme of Classification and Diagnosis Through the Simultaneous Use of Feature and State Spaces

The use of ontologies offers a distinct advantage: individual classification errors do not disrupt the overall reasoning process concerning the configuration or "portrait" of features that typify a given pathological process. This holds true whether the condition involves a single disease or a group of diseases sharing a common pathogenesis. A critical component of the disease classification process is the creation of a comprehensive bank of options for the development of pathological processes, which requires the use of artificial intelligence. It should be noted that the effectiveness of AI-based solutions in this context depends to some extent on the degree and validity of generalization of conditions, as well as on the peculiarities of structuring medical information [15].

Based on this rationale, a general framework for classification and diagnosis that concurrently employs both feature and state spaces may be outlined as follows:

1. Cleaning the data, evaluating the degree of inter-feature association, and assessing the informativeness, relevance and pertinence of features.
2. Establishing the relationships between distinct states, validating clusters of features corresponding to particular states, and evaluating the utility of naive Bayesian algorithms. Where appropriate, dynamic identification techniques and ontological modeling may also be incorporated.

The classification scheme shown in Figure 1 allows for the temporal inclusion of both the grouping of "similar" components into separate clusters and the reverse process, if necessary. Accordingly, disease classification refers to the systematic practice of distinguishing between different types of diseases by organizing conditions into separate categories based on certain criteria of similarity and difference [16]. The definition of a disease is influenced not only by the development

of medical science, but also by other factors, such as the capabilities of the relevant equipment, social problems, etc.

It is important to emphasize that the emergence of the paradigm of systems biology and systems medicine changes almost the entire strategy of disease classification, as well as diagnostic and prognostic processes. The new taxonomy requires the inclusion of the principles of systems science, namely, "complex systems, especially in biology and medicine, consist of dynamic, adaptive subsystems. These subsystems are controlled by competing communication channels and resulting emergent properties. Integration of various spatial and temporal modeling techniques is essential to refine and extend this new taxonomy [17, 18]. Systems biology and systems medicine are the means to consider the disease as a process and a reaction, the interface of triggering agents and the ongoing adaptation of the organism [19, 20]. From the point of view of systems medicine, the classification of diseases should take into account this expanded concept, giving the disease itself the characteristics of the involvement of systemic biology. In other words, disease manifestations are seen as the result of a continuous interaction between internal physiological mechanisms and external environmental factors. Accordingly, classification algorithms should be designed with the principles of scalability and communication in mind, reflecting the complex and interconnected nature of biological systems.

## CONCLUSIONS

1. It is proposed that when using artificial intelligence in healthcare for informed decision-making, two spaces - states and feature - should be used simultaneously.
2. The efficiency of classification can be significantly increased by adopting the principles of dynamic classification.
3. The effectiveness of using AI in healthcare depends to some extent on the level of generalization and the specifics of structuring medical information.
4. A critical and promising component of AI implementation is the development of a subject area ontology, which serves as a fundamental element at the preparatory stage of AI integration into healthcare.

## REFERENCES


1. Aldhyani THH, Alshebami AS, Alzahrani MY. Soft Clustering for Enhancing the Diagnosis of Chronic Diseases over Machine Learning Algorithms. J Healthc Eng. 2020;2020:4984967. doi: 10.1155/2020/4984967. DOI
2. Alfian G, Syafrudin M, Ijaz MF et al. A personalized healthcare monitoring system for diabetic patients by utilizing BLE-based sensors and real-time data processing. Sensors. 2018;18(7):2183. doi:10.3390/s18072183. DOI
3. Kumar Y, Koul A, Singla R et al. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. J Ambient Intell Humaniz Comput. 2023;14(7):8459-8486. doi: 10.1007/s12652-021-03612-z. DOI
4. Khanna M, Singh LK, Shrivastava K et al. An enhanced and efficient approach for feature selection for chronic human disease prediction: A breast cancer study. Heliyon. 2024;10(5):e26799. doi: 10.1016/j.heliyon.2024.e26799. DOI
5. Lyre H. The State Space of Artificial Intelligence. Minds & Machines 2020;30:325–347. doi:10.1007/s11023-020-09538-3. DOI
6. Mintser OP, Lukyanov E. Vikoristannya shtuchnogo intelektu na osnovi printsipiv samokontrolyu ta perehresnogo kontrolyu rishen, scho priymayutsya v biologiyi ta meditsini [The use of artificial intelligence based on the principles of self-control and cross-control of decisions made in biology and medicine.] Systems and means of artificial intelligence: abstracts of the International Scientific Conference «Artificial Intelligence: Achievements, Challenges and Risks.» Kyiv: IPAI «Science and Education», 15-16.03.2024. 2024, p.154-159. (Ukrainian)
7. Ledley RS. Syntax-directed concept analysis in the reasoning foundations of medical diagnosis. Computers in Biology and Medicine. 1973;3(2):89-99. doi:10.1016/0010-4825(73)90054-1. DOI
8. Ledley RS, Lusted LB. Reasoning Foundations of Medical Diagnosis. Science.1959;130(3366):9– 21. doi:10.1126/science.130.3366.9. DOI
9. Lovell BC, Bradley AP. The multiscale classifier. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996;18(2):124–137. doi:10.1109/34.481538. DOI
10. Krittanawong C, Zhang H, Wang Z et al. Artificial Intelligence in Precision Cardiovascular Medicine. J Am Coll Cardiol. 2017;69(21):2657-2664. doi: 10.1016/j.jacc.2017.03.571. DOI
11. Wang G-G, Deb S, Coelho LdS. Elephant Herding Optimization, 2015 3rd International Symposium on Computational and Business Intelligence (ISCBI), Bali, Indonesia. 2015, pp.1 – 5. doi: 10.1109/ISCBI.2015.8. DOI
12. Nilashi M, Ibrahim O, Ahmadi H et al. A knowledge-based system for breast cancer classification using fuzzy logic method. Telematics Inf. 2017;34 (4):133-144.
13. Prashad CD. State-space modelling for infectious disease surveillance data: Dynamic regression and covariance analysis. Infectious Disease Modelling. 2025;10(2):591-627. doi:10.1016/j.idm.2024.12.005. DOI

14. Chowell G. Fitting dynamic models to epidemic outbreaks with quantified uncertainty: A primer for parameter uncertainty, identifiability, and forecasts. Infectious Disease Modelling, 2017;2(3):379-398. doi:10.1016/j.idm.2017.08.001. DOI
15. Jafari-Marandi R, Davarzani S, Gharibdousti MS et al. An optimum ANN-based breast cancer diagnosis: Bridging gaps between ANN learning and decision-making goals. Appl. Soft Comput. 2018;72:108-120. doi:10.1016/j.asoc.2018.07.060. DOI
16. Wadmann S. Disease classification: A framework for analysis of contemporary developments in precision medicine. SSM - Qualitative Research in Health. 2023;3:100217. doi:10.1016/j.ssmqr.2023.100217. DOI
17. Mintser OP, Zaliskiy VM. Sistemna biomeditsina. T.1 Kontseptualizatsiya [Systemic biomedicine.T.1 Conceptualization]. Kyiv. Shupyk National Medical Academy of Postgraduate Education. 2020, p. 490. (Ukrainian)
18. Berlin R, Gruen R, Best J. Systems Medicine Disease: Disease Classification and Scalability Beyond Networks and Boundary Conditions. Front Bioeng Biotechnol. 2018;6:112. doi: 10.3389/fbioe.2018.00112. DOI
19. Bechtel W. Using the hierarchy of biological ontologies to identify mechanisms in flat networks. Biol. Philos. 2017;32:627–649. doi:10.1007/s10539-017-9579-x. DOI
20. Green S, Serban M, Scholl R et al. Network analyses in systems biology: new strategies for dealing with biological complexity. Synthese 2017;195:1751–1777. doi:10.1007/s/11229-016-1307-6. DOI

## CONFLICT OF INTEREST

The Authors declare no conflict of interest

## CORRESPONDING AUTHOR

**Olga O. Sukhanova**
Shupyk National Healthcare University of Ukraine
9 Dorohozhytska St., 04112 Kyiv, Ukraine
e-mail: olgasukhan@gmail.com

## ORCID AND CONTRIBUTIONSHIP

Ozar P. Mintser: 0000-0002-7224-4886 A D E F
Larysa Yu. Babintseva: 0000-0003-2753-5489 D E F
Stanislav I. Mokhnachov: 0000-0002-3480-9188 B D
Olga O. Sukhanova: 0000-0003-1882-027X B D F

A – Work concept and design, B – Data collection and analysis, C – Responsibility for statistical analysis, D – Writing the article, E – Critical review, F – Final approval of the article

**RECEIVED:** 11.01.2025
**ACCEPTED:** 20.06.2025