DOI: 10.63550/ICEIP.2025.45.84.009

The Use of Kohonen's Neural Network in Clustering of Parasitological Micrographs

Natalia Verezubova [1] [1], Artur Chekulaev [1] [2], Olga Yakovleva [1] [3], Natalia Sakovich [2] [4]
and Natalia Petrakova [2] [5]

[1] *Moscow State Academy of Veterinary Medicine and Biotechnology named after K.I. Skryabin, Academician Skryabin St., 23, Moscow, 109472, Russia*

[2] *Bryansk State Agrarian University, Sovetskaya St. 2a, Bryansk Region 243365, Russia*

nverez@mail.ru , tur2205@yandex.ru , yakovleffo@yandex.ru , nasa2610@mail.ru , npetrakova71@mail.ru

Keywords : neural network, parasitology, clustering.

Abstract: This article examines in detail a promising direction for the application of a neural network such as the Kohonen self-organizing map (SOM) in the field of applied parasitology. It considers theoretical justification of using this type of artificial neural networks for analyzing parasitological samples, with an emphasis on their ability to effectively cluster without preliminary data labeling. Various clustering quality metrics typical for self-organizing maps are presented and critically analyzed, including topological error, quantization error. As part of the experimental part of the work, a comprehensive study was conducted, including the creation and preparation of a specialized data set consisting of micrographs of the biohelminth Trichinella Spiralis . To implement the tasks set, a Kohonen network with an optimized architecture was developed and modeled in Python using modern machine learning libraries. Based on the results obtained, reasonable conclusions were formulated about the performance of the developed model, its accuracy and efficiency .

1 INTRODUCTION

Clustering method is a key tool in many systems, used for both object identification and data compression. There are many different ways to implement the clustering algorithm. These systems are divided into top-down and bottom-up, based on mathematical calculations, and methods using artificial neural networks .

One such method is a neural network such as the Kohonen self-organizing map (Self-organizing map) (SOM), which is a variant of the implementation of the vector quantization idea.

Vector quantization is a way of representing data as a set of vectors, each of which represents the center of some group (cluster) of similar data. SOM thus allows one to find correspondences between data and identify patterns within large sets of information by grouping similar data into clusters.

The choice of clustering method depends on the specifics of the problem and the properties of the data. The Kohonen network is a powerful tool for problems where it is necessary to find hidden structures and patterns in data, especially when these patterns are complex and nonlinear .

This paper presents an innovative approach to parasitological data analysis based on the use of Kohonen self-organizing maps (SOMs). Until now, such models have not been used in applied parasitology, which is probably due to the lack of familiarity with the potential of this method. However, as will be demonstrated, the use of SOMs can revolutionize workflows in this field, significantly reducing time and resource costs. A key advantage is the automation of large-scale data analysis, which is especially relevant for modern

[1] https://orcid.org/0009-0003-4139-8094
[2] https://orcid.org/0009-0005-7424-3106
[3] https://orcid.org/0009-0003-5953-3020
[4] https://orcid.org/0000-0002-6154-4445
[5] https://orcid.org/000 0 -0002-2744-3157

DOI: 10.63550/ICEIP.2025.45.84.009

parasitological studies that generate a huge number of microscopic images and other types of information.

Self-organizing maps, a type of artificial neural network, have a unique ability to cluster data without pre-setting cluster parameters. This feature makes them extremely attractive for problems where a priori information about the structure of the data is limited or absent. This is especially important in parasitology, since the diversity of parasitic organisms and their life cycles is extremely large, and classification is often difficult due to morphological similarities between species or developmental stages. Similar methods have been successfully applied in other areas of biology, for example, to classify animal species based on morphological features, genetic data, or ecological characteristics. Successful application in related fields indicates the high potential of SOM in parasitology as well.

Within the framework of this study, it is planned to develop and test the Kohonen self-organizing map model, specially adapted for the analysis of parasitological micrographs. This will allow us to evaluate the efficiency of SOM in automatic image clustering, identification of various forms of parasites, and determination of relationships between morphological features of parasites and their ecological niches or genetic characteristics. A detailed analysis of the results will allow us to evaluate the accuracy, speed , and efficiency of the proposed approach in comparison with traditional methods used in parasitological studies. It is expected that the results of this study will have significant practical significance, opening up new opportunities for accelerating diagnostics, increasing the accuracy of parasite identification, and for in-depth study of evolutionary processes in parasitic populations. Moreover, the development of this model can serve as a basis for creating automated systems for analyzing parasitological data, significantly facilitating the work of researchers and increasing the efficiency of scientific research in this important area of biology.

2 MATERIALS AND METHODS

The main material of the study will be a dataset representing a set of images (microphotographs) of the parasite Trichinella Spiralis. The sample size consisted of four classes (depending on the morphology of the parasite) and numbering four hundred images, taking into account augmentation by rotation.

The research method will be clustering of the presented images using a self-organizing Kohonen map of hexagonal topology, modeled in Python, containing a total of twenty-five neurons.

The neural network was optimized using the Grind Search hyperparameter search method.

The process of testing the neural network hyperparameters was carried out over 500 iterations, and its effectiveness was carefully monitored using a comprehensive set of metrics. A key aspect of monitoring was the use of not only traditional clustering quality assessment metrics, such as the silhouette coefficient and the Davis-Boldin index, but also the inclusion of SOM architecture-specific metrics: quantization error and topographic error. Using these additional parameters allowed us to obtain a more complete picture of the quality of the formed clusters and the internal structure of the trained network.

The silhouette coefficient is known to be one of the widely used metrics for assessing the quality of clustering, as it allows one to simultaneously analyze the cohesion (compactness) of clusters and their separation . This indicator takes values in the range from -1 to 1, where high positive values indicate that objects within one cluster have a high degree of similarity to each other and are also well separated from objects belonging to other clusters. Negative values of the coefficient, on the contrary, indicate possible errors in clustering, for example, that objects may have been incorrectly assigned to clusters. Thus, the silhouette coefficient provides an intuitive assessment of the quality of clustering, but its interpretation may be limited in cases where clusters have a complex shape or varying density.

The Davis-Bouldin index, in turn, is a more sensitive indicator that also evaluates the quality of clustering, but does so taking into account both the compactness of the clusters and the degree of their separation. This index is calculated as the average ratio of the maximum value of the sum of intracluster variance to the intercluster distance, which allows for the balance between the density of clusters and their isolation. A low value of the Davis-Bouldin index indicates high clustering quality, since it indicates that the clusters are compact and well separated. However, this index may be less robust to noise and outliers in the data, which requires additional preprocessing before its application.

These classic metrics, such as the silhouette coefficient and the Davis-Bouldin index, provide a general overview of the clustering results and are universal tools for assessing the quality of data separation into groups. However, despite their wide applicability, they have certain limitations, especially

DOI: 10.63550/ICEIP.2025.45.84.009

when it comes to specific clustering methods, such as self-organizing maps (SOMs). For example, the SOM architecture based on the topological ordering of neurons and unsupervised learning produces clusters that may not match the traditional assumptions about the shape and density of clusters inherent in these metrics. As a result, the use of standard metrics may not fully reflect the features and advantages of SOMs, which requires the development or adaptation of specialized metrics that take into account the topological structure and properties of this model. For this reason, additional metrics directly related to the structure of the self-organizing map were included. Quantization error, for example, assesses the accuracy of the representation of the input data in the map space. It is calculated by measuring the Euclidean distance between each input vector and the weight vector of the neuron to which this input vector was assigned. Thus, a low quantization error indicates that the input data is well approximated by the neuron weights, and the network successfully captures the structure of the data. The smaller the quantization error, the better the network "compresses" the information while preserving important features of the data.

This metric is usually calculated using the formula:

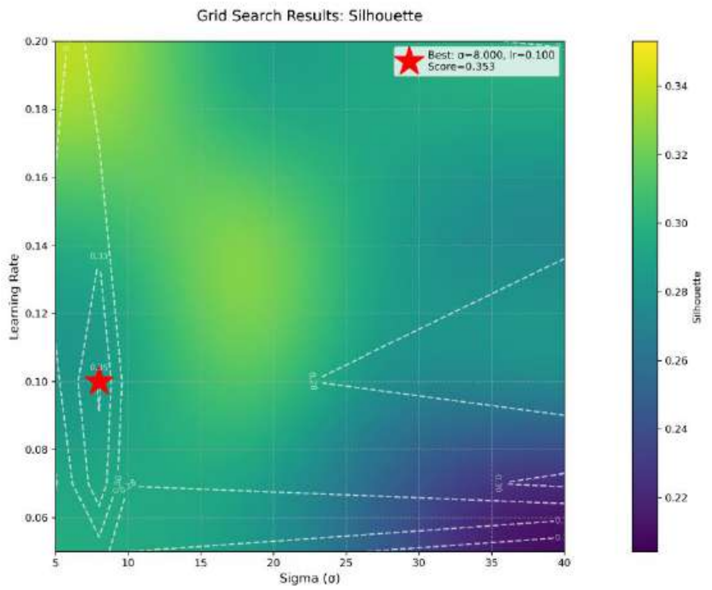
$$QE(\{\boldsymbol{m}_k\}, \mathbb{X}) = \frac{1}{N}\sum_{i=1}^{N}\left\|\boldsymbol{x}_i - \boldsymbol{m}_{b_i}\right\|_2$$

Topographic error, in turn, reflects the preservation of the topological properties of the original data in the structure of the self-organizing map. It measures how closely neurons corresponding to close input vectors are located in the map space. A high value of topographic error indicates a distortion of the topology, when close points in the original data space are mapped to distant neurons of the map, which indicates problems with the quality of training and data display. Ideally, the topographic error should be minimal, which means that spatial relationships between data are preserved in the network structure.

In combination, the use of the silhouette coefficient, Davis-Boldin index, quantization error and topographic error allowed us to perform a comprehensive analysis of the quality of training of the self-organizing map, taking into account both the general indicators of clustering quality and the specific characteristics of the internal structure of the network. Monitoring these parameters during 500 training iterations allowed us to optimize the process and obtain the best clustering results.

3 RESULTS AND THEIR DISCUSSION

As a result of the lattice hyperparameter search method, conducted using the parameters learning rate (learning rate) and sigma ( ), an iteration with a silhouette coefficient equal to 0.353 was selected. The heat map shown in Fig. 1 demonstrates the results of δhyperparameter tuning .



DOI: 10.63550/ICEIP.2025.45.84.009

Hyperparameter search graph and heatmap .

After the optimization was completed, training was carried out on one thousand iterations, the results of the topography and quantization errors are presented in Fig. 2.

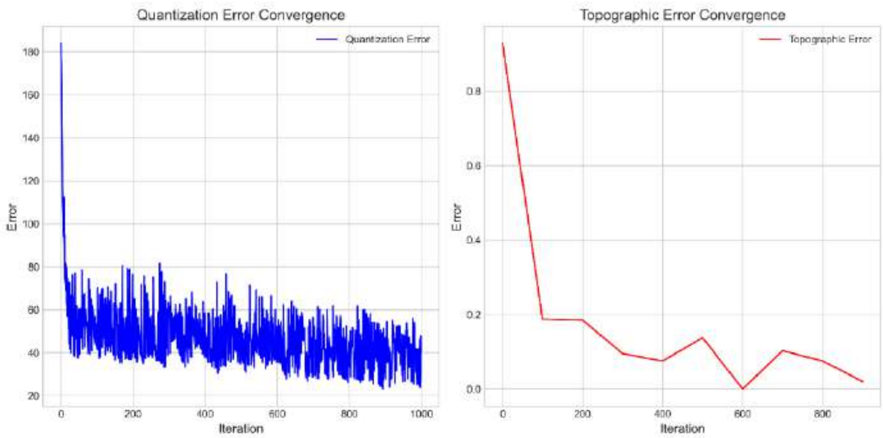


Figure 2: Topographic error and quantization error in dynamics.

Analysis of the training results of the neural network used to process helminth images reveals a number of interesting features that require detailed consideration. The key indicator of training efficiency is the reduction in topographic error. This reduction directly indicates the improvement of the self-organization of the Kohonen map (or similar self-organizing map) underlying the model. During the training process, the map neurons increasingly adapt to the input data, minimizing the difference between the input images and the activation of the “winning” neurons, i.e. those neurons that most accurately match the input data. Thus, the observed reduction in topographic error indicates an increasing “affinity” between the map neurons and the characteristics of the output data, which reflects an increasingly accurate representation of the input information by the neural network.

In parallel with this process, a decrease in the quantization error is observed with each training iteration. The quantization error is a measure of the distance between the original data (helminth images) and their projections on the Kohonen map. A decrease in this error illustrates the gradual approach of the projections to the original data, which indicates an increasingly accurate mapping of the data by the model. It is important to emphasize that the quantization error asymptotically approaches zero. This is a desirable result, since a complete zeroing of the error could indicate overtraining of the model - a situation when the network "remembers" the training sample, but poorly generalizes knowledge to new data. The fact of incomplete zeroing of the error confirms the adequacy of the training process and the ability of the model to effectively generalize information.

However, the analysis of the results is complicated by the peculiarities of data preparation. However, the number of clusters identified by the model does not coincide with the true number of clusters, which is due to the use of data augmentation techniques (artificially increasing the training sample by modifying existing images). Augmentation, although it improves the generalizing ability of the model, can distort the natural distribution of the data, leading to an inaccurate determination of the number of clusters.

In combination with the low value of the silhouette coefficient (a metric that evaluates the quality of clustering), this circumstance indicates a high degree of similarity and homogeneity of images in the training sample. Perhaps the initial diversity of helminth morphological features was not large enough, or augmentation led to a "blurring" of the boundaries between classes. For a more accurate analysis, it is necessary to conduct additional research, including studying the influence of augmentation parameters and using other methods for assessing the quality of clustering. The obtained results allow us to conclude that the model was trained effectively, but require a more detailed analysis taking into account the specifics of data preparation and processing.

As part of the study, an artificial neural network was developed and implemented, the functioning of which is based on the principles of self-organizing maps, known as Kohonen maps (Self-Organizing Maps , SOM). This approach, proposed by Finnish scientist Teuvo Kohonen , is a powerful machine learning tool focused on high-dimensional data clustering and visualization tasks. Unlike traditional clustering methods, SOM has the unique ability to preserve the topological properties of the input data, making it particularly effective for analyzing complex and high-dimensional data sets.

To conduct the experiment, a sample of microphotographs related to the field of parasitology was formed. It should be noted that the sample was relatively limited in volume, which is due to the specifics of the available data and the complexity of obtaining high-quality images of parasitic organisms. Nevertheless, despite the limited sample, the results of the experiment were very informative and allowed us to draw a number of important conclusions. During testing, an artificial neural network based on Kohonen maps demonstrated a high level of efficiency in the tasks of clustering microphotographs. Kohonen maps not only successfully grouped data according to their internal characteristics, but also showed satisfactory results in the tasks of identifying objects based on the clustering performed. This confirms their potential as a tool for automated image analysis in biomedical research.

The obtained results allow us to state that Kohonen self-organizing maps have significant advantages over traditional data analysis methods. First, they are able to efficiently process high-dimensional data while preserving their topological structure, which is especially important when analyzing complex biological images. Second, SOMs demonstrate a high degree of adaptability, which allows them to be used to work with data of various natures and structures. Third, the clustering results obtained using SOMs can be easily interpreted, which makes them a convenient tool for researchers working in the field of parasitology.

4 CONCLUSION

Thus, it can be confidently stated that the use of Kohonen self-organizing maps in the field of parasitology is a promising and promising direction. This opens up new horizons for further research aimed at improving the methods of automated analysis of microphotographs, as well as for the practical use of such technologies in the diagnosis and classification of parasitic organisms. It is important to emphasize that the introduction of such technologies can significantly simplify and speed up the process of analyzing microphotographs, which is especially important in conditions of limited resources and the need for prompt processing of large volumes of data. Automation of image analysis using SOM allows minimizing the human factor, reducing the likelihood of errors and increasing the accuracy of diagnosis.

In addition, the use of self-organizing maps can contribute to a deeper understanding of the structure and characteristics of parasitic organisms, which in turn opens up opportunities for the development of new methods for the treatment and prevention of parasitic diseases. For example, SOMs can be used to identify hidden patterns in data that are not obvious using traditional analysis methods. This can become the basis for the creation of more accurate diagnostic systems capable of identifying parasites at early stages of their development, which is especially important for timely treatment.

It should also be noted that the successful application of SOM in this study highlights the need to further explore the potential of this technology in biomedical sciences. In the future, Kohonen maps can be integrated into more complex artificial intelligence systems, which will allow the creation of comprehensive solutions for the analysis of biomedical data. Such systems can become an indispensable tool for researchers and specialists in the field of parasitology, providing more accurate, faster and more reliable results.

In conclusion, the use of Kohonen self-organizing maps in parasite research represents a significant step forward in the development of parasitology. This technology not only demonstrates high efficiency in clustering and object identification tasks, but also opens up new opportunities for automation and optimization of data analysis processes.

REFERENCES

Kudinova , V. A. , 2023. Klasterizaciya i eyo methods . Hydroelectric power stations v XXI century : Sbornik materials X All-round scientific - practical conferences young scholars , specialists , aspirants , students i shkol'nikov , Sayanogorsk ; CHeremushki , 24–28 April 2023 . Sayano-SHushenskij filial SFU. Sayanogorsk ; CHeremushki : Sayano-SHushenskij filial SFU, 2023. S. 372-376 .

Akume , D., Weber, G. W., 2002. Cluster algorithms: theory and methods. *Computational Technologies.* Vol. 7, No. 1. P. 15-27.

DOI: 10.63550/ICEIP.2025.45.84.009

Tatarnikova , TM, Poymanova , ED, 2019. Algorithms for placing files in tiered storage using Kohonen map. CEUR Workshop Proceedings: ISTMC 2019. Selected Papers of the 4th All-Russian Scientific and Practical Conference with International Participation "Information Systems and Technologies in Modeling and Control", Yalta, Crimea, 21–23 maya 2019 . Vol. 2522. Yalta, Crimea: CEUR, 2019. P. 193-202.

Kohonen , T., 1998. The self-organizing map. *Neurocomputing* . Vol. 21, No. 1-3. – P. 1-6.

Mavroudi , S., Papadimitriou, S., Bezerianos , A., 2002. Gene expression data analysis with a dynamically extended self-organized map that exploits class information. *Bioinformatics.* Vol. 18, No. 11. P. 1446.

Minikaev , RF, YAkupov , ZYA, 2010. Nejronnye seti i cards Kohonena . *ZHurnal Srednevolzhskogo mathematical society* . V. 12, No. 3. – S. 86-95.

Using the quantization error from Self-Organized Map (SOM) output for detecting critical variability in large bodies of image time series in less than a minute. arXiv.org URL: https://arxiv.org/abs/1710.10648 (data obrashcheniya : 03/09/2025).

Backer, E., Jain, AK, 1981. A clustering performance measure based on fuzzy set decomposition. IEEE Trans Pattern Anal Mach Intell 1:66–75.

A Survey and Implementation of Performance Metrics for Self-Organized Maps. arXiv.org URL: https://arxiv.org/abs/2011.05847 (data obrashcheniya : 03/09/2025).

Evaluating Self-Organizing Map Quality Measures as Convergence Criteria. semanticscholar.org URL: https://api.semanticscholar.org/CorpusID:53065444 (data obrashcheniya : 03/09/2025).