International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

# A Novel Model for Image Classification based on Persistent Homology

**Petar Sekuloski[1], Vesna Dimitrievska Ristovska[2], Vassil Grozdanov[3]**

[1, 2]Faculty of Computer Science and Engineering, "Ss. Cyril and Methodius" University, Skopje, Macedonia
[1]*petar. sekuloski[at]finki.ukim. mk,*
[2]*vesna.dimitrievska. ristovska[at]finki.ukim*

[3]Department of Mathematics, Faculty of Mathematics and Natural Sciences, South - West University NeophitRilski, Blagoevgrad, Bulgaria
*vassgrozdanov[at]yahoo.com*

**Abstract:** ***Topological Data Analysis (TDA) is relatively new field of Applied Mathematics that emerged rapidly last years. The main tool of Topological Data Analysis is Persistent Homology. Persistent Homology tracks the topological features of datasets. In this paper we will introduced a novel model for image classification based on Persistent Homology. In the experimental part we the introduced new novel on real medical dataset. Using this model, the classification was improved.***

**Keywords:** Topological Data Analysis, Persistent Homology, Machine Learning, Computational Topology, Image Classification

## 1. Introduction

Image classification is a topic has high tendency in both research areas machine learning and computer vision. Everyday usage of technology such as high - quality cameras, social media platforms, social networks and etc. has provided many digital images that can be used in machine learning models. Also, technological developments resulted with big datasets of digital images in many areas such as health care, astrophysics, biology, geodesy and others. In this paper, we introduce a new novel model for classification of digital images based on Persistent Homology and deep neural networks.

Homology is a mathematical concept which associates sequences of algebraic objects with topological spaces. One way to study a topological space is to find and compute its homology groups. The motivation behind defining homology groups was that two shapes can be distinguished by examining their holes. For example, a disk is different from a circle, or a disk is not a circle, because the disk is solid while the circle has a hole through it. Homology groups are set of invariants of a topological space. Homology groups characterize the topological space. The main idea of Persistent Homology is to track the topological characteristics of a reconstructed space from a dataset.

The advantages of using Persistent Homology are tracking the global signatures of the datasets such as robustness and invariance of the topological signatures. Topological signatures are more resistant to local deformations and computations of these signatures do not depend on the scale of data.

## 2. Literature Survey

Although TDA is a relatively new field, in the last 4 years there have been a large number of scientific papers involving topological data analysis and it has been successfully applied in a number of fields. Some of the areas covered by the application of TDA are: gene expression, neural network analysis, chemoinformatics, time series prediction, cancer detection, cyber security, eco - informatics, natural language processing, sound processing, face recognition, analysis and time series prediction, stability of dynamical systems, image segmentation, sensor networks, complex networks, banking, sensor networks, noise detection, signal processing, bioinformatics, and many others [1 - 10].

On this occasion, we will give several successful examples.

Persistent Homology have been used with neural networks for graph classification problems in [6]. In [7], the similarity of the neural connections of the brains of twins is estimated using TDA. In [8], persistent homology are used for brain imaging. In [9], exceptionally good results were obtained for detecting a subset of breast cancer.

In [10] persistent images are used for trajectory - related classification of a dynamic system, as unique attributes, and in combination with SVM give good results.

## 3. Model

In this work it is applied Persistent Homology on digital images. CW - complexes are generalization of simplicial complexes that allow cells that are not necessarily simplices, homeomorphic to balls or open discs [10, 11]. For example, cubes instead of tetrahedra. In this work it is used regular CW - complexes. Because of the nature of the data, images are matrices, some grid structures, will use cell complexes instead of simplicial complexes.

**Definition 1.** A filtered cell complex is $(X, F)$ is a cell complex $X$ together with a monotonic function $f: X \to \mathbb{R}$. A linear ordering $\sigma_0, \sigma_1, \sigma_2, \dots, \sigma_n$ of the cells in $X$, such that $\sigma_i \preccurlyeq \sigma_j$ implies $i \leq j$, is compatible with the function $f$ when

$$f(\sigma_0) \leq f(\sigma_1) \leq f(\sigma_2) \leq \cdots \leq f(\sigma_n)$$

The novel introduced in this model includes construction of regular cell complexes of the images, the computation persistent images from the constructed complexes. In the *Figure 1*, there is an example of construction of cell complex from a digital image together with persistent homology More on construction of cell complexes from digital images and computing Persistent Homology of cell complexes can be found in [11].

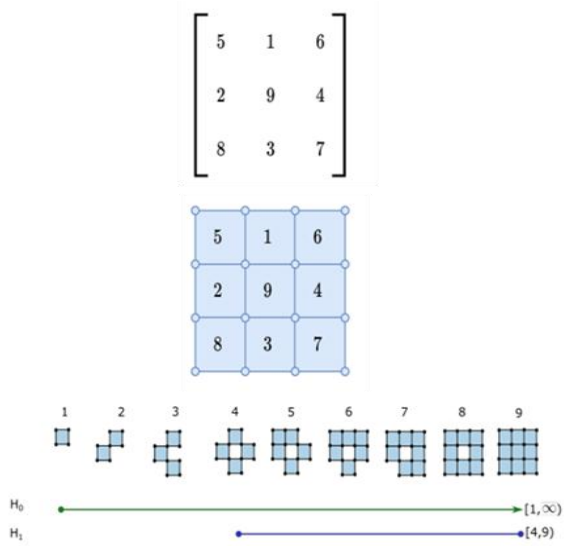


**Figure 1:** An example of constructing cell complex from two dimensional image together with persistent homology.

The novel model for image classification based on Persistent Homology introduced in this paper has four phases: Preprocessing the Data, Computation of Topological Features, Concatenating the Data and Classification. The model is shown on Figure 2.

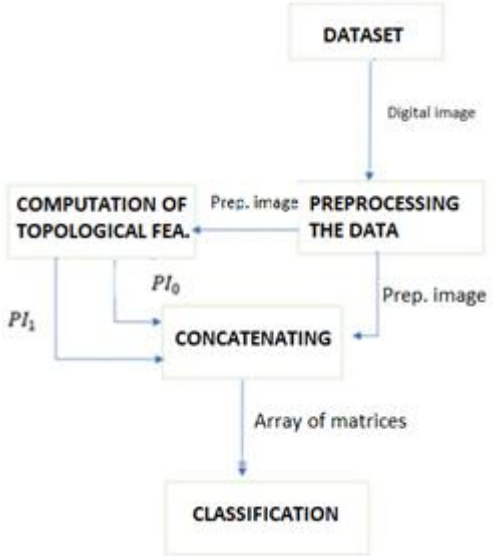


**Figure 2:** A model for image classification based on Persistent Homology

The first phase of the model – Preprocessing of the data depend on the dataset that is used for a classification task. In the following section of this paper we will discuss more on the preprocessing of the date used in our experiments.

The second phase of the model is the most important for Persistent Homology. First, for each of the images in the dataset, a cubic complex is constructed, with a V - construction. Then, homology is calculated, on the already obtained complex from the image. Once the homology is computed, the Persistent Images can be computed. More on Persistence Diagrams can be found in [10]. This phase is shown in Figure 2.

Persistent Image is bound only to the persistent groups of a certain dimension, and since we have two - dimensional images, for the images we worked with, the dimension can be 0 or 1. Therefore, from the persistent diagram, 2 persistent images are obtained for each of the digital images from the dataset. And in fact, this is the main purpose of this work, how the extracted topological attributes from the images will affect the classification process.

Next, the Persistent images, in Figure 3, denote as$PI_0$, which corresponds to the persistent groups whose dimension is 0, and $PI_0$, which corresponds to the persistent groups whose dimension is 1, together with the original images enter the phase - Concatenating of data. In this phase, from the original image, and the two persistent images, we create a sequence of three matrices, that is, we create a three-channel representation of each of the images that we have in the data set. In the first place in the sequence, or in other words on the first channel, is the original image, and on the second and third positions, i.e. channel, are $PI_0$and$PI_1$.

We will make two experiments, first a classification task in which topological features are not included, and second, we will perform the classification task with the new model described in this section. As a classifier in both experiments the same neural network is used.

## 4. Experiments and Results

**a) Br35H:: Brain Tumor Detection Dataset**

Br35H:: Brain Tumor Detection [16] is a public dataset containing 3000 patient images obtained by magnetic resonance imaging of the head region. It is intended for brain tumor detection. The images are divided into two classes: 1500 of the images belong to the "No" class, the class indicating that the images are obtained from patients which do not have a tumor, and the "Yes" class, which indicates that the images are obtained from patients who have a tumor.

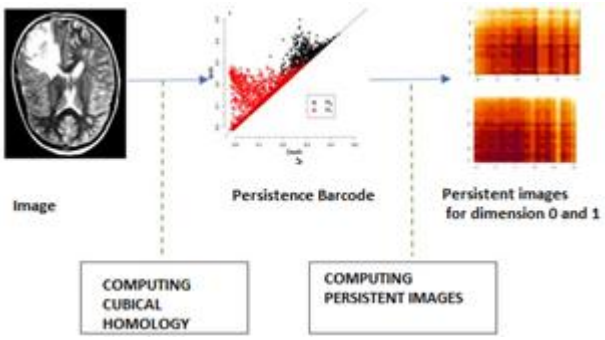


**Figure 3:** Computation of Topological Features

Why did we choose this dataset? - Because the detection of a tumor requires a specialist opinion, especially when it is in an early stage and it is more difficult to detect it. The need for a rapid response in some types of tumors is crucial for patients' lives. Also, making a diagnosis itself is a challenge in this area. The images in the dataset are grayscale images with different dimension and pixel value from 0 to 255. We also resized all the images to 128×128. The set is balanced, having the same number of images from both classes.

**b) Preprocessing and computing Persistent Images**

The data is not divided into a test set and a training set, so the first step with this set from the preprocessing was to split it into two sets, one of which will be used for training and the second for testing. We constructed the test set by randomly selecting 200 images from one class, and 200 images from the other class. We obtained a test set containing 400 images of both classes. While for the training set we are left with 1300 images for each of the classes. We kept the training set balanced. During preprocessing we scaled all the images so that the pixel value will be in the range from 0 to 1.

After this phase, we moved on to merging the persistent images with the original image. We have created a three member array whose elements are matrices of dimension 128×128. The first member of the array is the original image, while the second members are the persistent images obtained in the previous stage. Let's emphasize that we do this transformation on both the training subset and the testing subset.

**c) Classification**

First, we will describe the classifier that we used in our experiments. We used neural network for the classification tasks. The first layer in the network, layer_flatten, transforms the input from a three - element array, whose elements are matrices of dimension 128×128, into an array by concatenating the rows of the three matrices into one row, side by side. In this model, in which the attributes obtained from the Computation of Topological Features phase are not used, a three - channel image is given as input and all three channels have the original image. This is how the array of matrices is formed. In the hidden layer of the network there are two dense layers. They are fully connected neuronal layers with 128 neurons each. And in the last layer, which contains 10 neurons, as many classes as we have for classification, the activation function "softmax" determines the probabilities to which class the image belongs. "adam" optimization and "sparse_categorical_crossentropy" as loss function were used.

On Figure 4you can see the accuracy value and loss function across epochs during training. We stopped training at the 200th epoch, as both the accuracy and the loss function converged after the 70th epoch. We used a neural network because it is most suitable for working with images. On the test set we got an accuracy of 0.718, and a loss of 0.4999. The confusion matrix is given in Table 1.

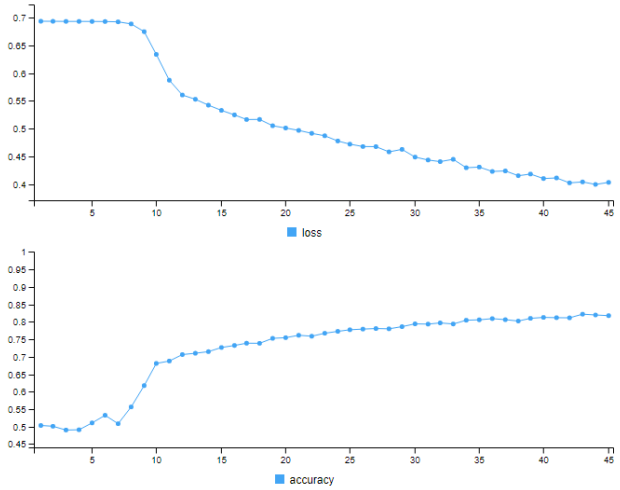


**Figure 4:** Accuracy and loss of the classification task without topological features

The confusion matrix is given in Table 1. For evaluating, we used precision, recall and f1 - score as metrics except accuracy. The values of these metrics are given on Table 2.

**Table 1:** Confusion matrix of the classification task without topological features

| Actual/ Predicted | 0 | 1 |
|---|---|---|
| 0 | 143 | 54 |
| 1 | 44 | 156 |

**Table 2:** Results of the evaluation of the classification task without topological features

| Class | Precision | Recall | F1 - Score |
|---|---|---|---|
| 0 | 0.7647 | 0.7258 | 0.744 |
| 1 | 0.742 | 0.78 | 0.76 |
| Average | 0.753 | 0.75294 | 0.75288 |

To test the novel model introduced in this paper, we created a three - member array of 128×128 matrices in the concatenation phase. The first member of this array is the original image, the second member is the persistent image corresponding to the persistent homology groups with dimension 0 and the third matrix is the persistent image corresponding to dimension 1. Before we choose the standard deviation when creating the persistent images to be 0.00001 we did some experiments, the purpose of which was to select the most appropriate value of this parameter. We used the same neural network architecture as we used in previous experiment.

We stopped training at the 200th epoch, as both the accuracy and the loss function converged after the 40th epoch. We used a neural network because it is most suitable for working with images. On the test set we got an accuracy of 0.924, and a loss of 0.38. The results for the other metrics are shown in Figure 5.

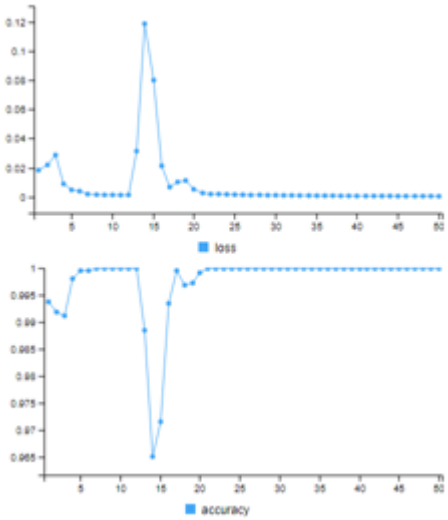


**Figure 5:** Accuracy and loss of the classification task with topological features

The confusion matrix is given in Table 1. For evaluating, we used precision, recall and f1 - Score as metrics except accuracy, as we used in the previous experiment. The values of these metrics are given on Table 2.

**Table 3:** Confusion matrix of the classification task with topological features

| Actual/ Predicted | 0 | 1 |
|---|---|---|
| 0 | 179 | 19 |
| 1 | 11 | 189 |

**Table 4:** Results of the evaluation of the classification task with topological features

| Class | Precision | Recall | F1 - Score |
|---|---|---|---|
| 0 | 0.9417 | 0.9035 | 0.9223 |
| 1 | 0.9086 | 0.9450 | 0.9264 |
| AVERAGE | 0.9252 | 0.9242 | 0.9423 |

In Table 5, a summary analysis of the improvement of the classification using TDA compared to the classification without using TDA, for the same classifier, is given.

**Table 5:** Summary results of the testing process of the two classification tasks

| | Precision | Recall | F1 - Score |
|---|---|---|---|
| Without PH | 0.718 | 0.753 | 0.753 |
| With PH | 0.924 | 0.925 | 0.942 |
| Difference | 0.206 | 0.172 | 0.189 |
| Percentage of Improvement | 28.69 | 22.84 | 25.1 |

In this data set, if we compare the accuracy using topological features which is 0.924, against the accuracy of the classifier without using topological features which is 0.718, we can say that there is an increase of 28.7%.

Also if we compare the rest of the obtained metrics, they are all significantly better with our model proposed in chapter 4, and the same improvements expressed in percentages are: 28.7% for precision, 22.8% for response, 24% for loss function, 25.1% for the parameter f1 - Score.

In summary: all the results obtained with our new proposed model are 20% to 30% better, compared to the other model (without TDA). Knowing that in image classification, neural networks have superiority over other classifiers, that's why we used them.

## 5. Conclusion

To classify images from this problem, there are quite complex neural architectures that mostly use convolutional networks and transfer learning and obtain comparable results with this model, which is simple and includes a neural network with two hidden layers. During our experiments, the Calculation of topological features phase takes the most time, but even for such a dataset it is measured in minutes, performed on not very powerful hardware machines that do not include GPU and similar technologies.

For the dataset used in the experiment, which is a dataset of collected MRI images, and for which the best classifiers, neural networks, have complex architectures, the proposed model that includes topological features gave significantly better results than the model that does not include topological features. The improvement is 28.7% in accuracy, and in the other metrics that we used is from 20% to 30%, although it is a very simple model that does not require a lot of processing power.

## 6. Further Work

In the future, the influence of topological features can be investigated in some more complex classifiers or a more complex model can be built that will include them. Also, during the training process, a loss function could be defined based on topological characteristics.

In addition, it is possible to develop and study the input parameters of selected algorithms in the field of TDA, in order to consider the influence of the choice of parameter values on the obtained results for specific datasets.

**Acknowledgement**
The research presented in this paper is partly supported by the Faculty of Computer Science and Engineering, at the Ss. Cyril and Methodius University in Skopje.

## References

[1] M. Pirashvili, L. Steinberg, F. G. Guillamon, M. Niranjan, J. G. Frey, J. Brodzki, “Improved understanding of aqueous solubility modeling through topological data analysis”, Journal of Cheminformatics, (2018)
[2] J. Nicponski, J. H. Jung, “Topological data analysis of vascular disease: A theoretical framework”, BioRxiv, (2019)
[3] P. Sekuloski, V. D. Ristovska, Application of Persistent Homology on Bio - Medical Dataset – A Case Study, Mathematical Modeling - Proceedings, III International Scientific Conference, Bulgaria, (2019)
[4] P. Sekuloski, V. D. Ristovska, Mapper „Algorithm and It’s Applications, Mathematical Modeling – Proceedings “, III International Scientific Conference, Bulgaria, (2019)

[5] V. de Silva, R. Ghrist, Coordinate - free coverage in sensor networks with controlled boundaries via homology, International Journal of Robotics Research, vol 25, 1205 – 1222, (2006)
[6] Don, A. P. H.; Peters, J. F.; Ramanna, S.; Tozzi, A. Topological View of Flows Inside the BOLD Spontaneous Activity of the Human Brain. Front. Comput. Neurosci.2020, 14, 34.
[7] Carrière, M.; Chazal, F.; Ike, Y.; Lacombe, T.; Royer, M.; Umeda, Y. Perslay: A neural network layer for persistence diagrams and new graph topological signatures. In Proceedings of the International Conference on Artificial Intelligence and Statistics (PMLR), Online, 26–28 August 2020; pp.2786–2796
[8] Chung, M. K.; Lee, H.; DiChristofano, A.; Ombao, H.; Solo, V. Exact topological inference of the resting - state brain networks in twins. Netw. Neurosci.2019, 3, 674–694. [CrossRef] Axioms 2022
[9] M. Nicolau, A. J. Levine, G. Carlsson, Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival
[10] H. Adams, T. Emerson, M. Kirby, R. Neville, Ch. Peterson, P. Shipman, S. Chepushtanova, E. Hanson, F. Motta, L. Ziegelmeier, Persistence Images: A Stable Vector Representation of Persistent Homology", vol.18, p.1−35, (2017)
[11] I. Hatcher, Alebraic Topology, Cambridge University Press, (2002)
[12] J. R. Munkers, Topology, Upper Saddle River: Prentice Hall, Vol.2, (2000)
[13] B. Bleile, A. Garin, T. Hesis, K. Maggs, V. Robins, "The Persistent Homology of Dual Digital Images Constructions", arXiv: 2102.11397, (2021)
[14] A. Zomorodian. G. Carlsson. Computing persistent homology, Discrete & Computational Geometry, vol.2, p.249–274, (2005)
[15] H. Adams, T. Emerson, M. Kirby, R. Neville, Ch. Peterson, P. Shipman, S. Chepushtanova, E. Hanson, F. Motta, L. Ziegelmeier, Persistence Images: A Stable Vector Representation of Persistent Homology", vol.18, p.1−35, (2017)
[16] https: //www.kaggle. com/datasets/ahmedhamada0/brain - tumor - detection, пристапено на 15.08.2022, last accessed on 27.07.2022

## Author Profile

**Petar Sekuloski** is a teaching and research assistant at the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, in Skopje. Currently he is a M. Sc. student at the Faculty of Computer Science and Engineering. His research includes Topological Data Analysis, Machine learning and application of some method from algebraic topology to Computer Science.

**Vesna Dimitrievska Ristovska** is an associate professor at the Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, in Skopje. Her research interests include uniform distribution of sequences, numerical analysis, optimization methods, and topological data analysis.

**Vassil Grozdanov** is a professor Department of Mathematics, Faculty of Mathematics and Natural Sciences, South - West University Neophit Rilski, Blagoevgrad, Bulgaria. His research interests include uniform distribution of sequences, numerical analysis and optimization methods.