

Topological Data Analysis for COVID-19 Classification from Lung CT-Scan Images

Jeki Saputra^{1*}, Intan Muchtadi-Alamsyah²

¹Master's Program in Mathematics, Institut Teknologi Bandung, Indonesia

²Algebra Research Group, Faculty of Mathematics and Natural Sciences, Institut Teknologi Bandung, Indonesia

Abstract. Topological Data Analysis (TDA) has emerged as a promising framework for capturing structural patterns in medical images; however, existing approaches for COVID-19 CT classification, particularly those based on Riemannian representations of persistence diagrams, often involve complex non-Euclidean computations that limit scalability and integration with standard machine learning models. In this paper, we propose a TDA-based classification framework that utilizes persistence landscapes as a computationally efficient Euclidean representation of persistent homology. Each CT image is transformed into a Feature Point Cloud via an intensity-based covering and spatial clustering procedure, from which topological features (H_0 , H_1 , H_2) are extracted using Vietoris–Rips filtration, along with an additional descriptor obtained from lower-star filtration. These features are vectorized into a fixed-length representation and used for classification. Experiments on a balanced dataset of 1000 CT images with a 70:20:10 train/validation/test split demonstrate that the proposed method achieves accuracies of 96% using Support Vector Machines and 98% using Random Forest. An ablation study further reveals that the H_0 feature consistently provides the strongest discriminative performance, indicating that connectivity-based structures are more robust and informative than higher-dimensional topological features for distinguishing COVID-19 CT images, while maintaining computational efficiency and compatibility with classical machine learning approaches.

Key words and Phrases: persistence homology, persistence landscapes, COVID-19, CT-scan.

1. INTRODUCTION

Computed tomography (CT) imaging plays an important role in the diagnosis and monitoring of COVID-19 infections. Radiological studies have shown that

*Corresponding author: 20123305@mahasiswa.itb.ac.id

2020 Mathematics Subject Classification: 55N31, 68T10, 92C55

Received: 11-01-2026, accepted: 02-05-2026.

COVID-19 CT scans often exhibit characteristic patterns such as bilateral lung involvement and ground-glass opacities (GGOs). These structural patterns motivate the development of automated methods for analyzing CT images to assist clinical diagnosis.

Machine learning and deep learning approaches have been widely used for medical image classification. Convolutional Neural Networks (CNNs), including architectures such as EfficientNet and ResNet, have demonstrated strong performance in image recognition tasks [1, 2]. However, such models typically require large training datasets and often lack interpretability, making it difficult to understand the structural patterns that drive classification decisions.

Topological Data Analysis (TDA) offers an alternative framework for analyzing complex data by extracting geometric and topological features [3]. Persistent homology, one of the main tools in TDA, captures multiscale structures such as connected components, loops, and cavities [4]. These features provide interpretable descriptors that explicitly encode structural information in the data.

Previous work by Iqbal et al. [5] applied TDA to COVID-19 CT image classification by constructing feature point clouds and representing persistence diagrams using a Riemannian framework. While effective, this approach introduces additional computational complexity and relies on non-Euclidean statistical tools, which limits scalability and complicates integration with standard machine learning models.

This limitation motivates the need for a simpler and more computationally efficient representation of persistence diagrams. In this paper, we adopt persistence landscapes as a vectorization method that embeds persistence diagrams into a Euclidean space, enabling direct compatibility with classical machine learning algorithms while preserving essential topological information.

Following the geometric framework of [5], we construct Feature Point Clouds from CT images and compute persistent homology using Vietoris–Rips filtrations together with a lower-star filtration on the image grid. The resulting topological features are then vectorized and used for classification via Support Vector Machines and Random Forest classifiers.

Furthermore, we investigate the contribution of individual homological features through an ablation study. The experimental results show that the H_0 feature consistently provides the strongest discriminative performance, suggesting that connectivity-based structures are more robust and informative than higher-dimensional topological features for distinguishing COVID-19 CT images.

2. PRELIMINARIES

This section reviews the main concepts of topological data analysis and persistent homology used in this paper, with emphasis on their role in constructing features from image data. For further background, we refer to [3, 4].

2.1. Topological Data Analysis and Persistent Homology.

Topological Data Analysis (TDA) studies the qualitative structure of data using tools from algebraic topology. In many applications, geometric and structural properties such as connectivity, loops, and higher-dimensional cavities provide important information about the underlying data.

Persistent homology is a fundamental tool in TDA that captures such topological features across multiple scales. By tracking how these features appear and disappear along a filtration, persistent homology provides a multiscale summary of the data that is stable under small perturbations.

2.2. Simplices and Simplicial Complexes.

A k -*simplex* is the convex hull of $k + 1$ affinely independent points $v_0, \dots, v_k$, denoted by $[v_0, \dots, v_k]$. Examples include a point ($k = 0$), a line segment ($k = 1$), a filled triangle ($k = 2$), and a filled tetrahedron ($k = 3$).

A *simplicial complex* K is a finite collection of simplices such that every face of a simplex in K also belongs to K , and the intersection of any two simplices is either empty or a common face. Simplicial complexes provide a combinatorial representation of geometric structures and serve as the foundation for computing topological features.

2.3. Homology.

Let K be a simplicial complex with coefficients in $\mathbb{Z}_2$. The k -th homology group is defined as

$$H_k(K) = \ker(\partial_k) / \text{im}(\partial_{k+1}),$$

and its dimension $\beta_k = \dim H_k(K)$ is called the k -th Betti number. Intuitively, β_0 counts connected components, β_1 counts loops, and β_2 counts voids. These quantities provide compact summaries of the structural properties of data.

2.4. Filtrations and Persistent Homology.

A *filtration* is a nested sequence of simplicial complexes

$$K_0 \subseteq K_1 \subseteq \dots \subseteq K_T.$$

Applying homology to each K_t yields a sequence of vector spaces in which topological features appear and disappear. A feature is said to be *born* at a certain scale and *dies* at a later scale, and its lifetime is called persistence. Features with longer persistence are typically considered more significant.

2.5. Persistence Diagrams.

For a fixed k , the persistence diagram Dgm_k is a multiset of points $(b, d) \in \mathbb{R}^2$ with $b < d$, where b and d denote the birth and death of k -dimensional features. Points far from the diagonal $b = d$ correspond to persistent (and thus more significant) features, while points near the diagonal are often regarded as noise.

2.6. Persistence Landscapes.

Persistence diagrams are not vector-valued objects, which makes them difficult to use directly in machine learning. Persistence landscapes provide a functional embedding that maps persistence diagrams into a vector space [6].

For each point (b, d) in a persistence diagram, define the tent function

$$\lambda_{(b,d)}(t) = \max\{0, \min(t - b, d - t)\}.$$

The k -th persistence landscape $\lambda_k(t)$ is defined as the k -th largest value among all such functions at time t . The collection $\{\lambda_k\}_{k \geq 1}$ is called the persistence landscape.

In practice, the landscape is truncated to a finite number of layers and each function is sampled on a fixed grid to obtain numerical feature vectors. This representation is particularly important in our framework, as it enables persistence diagrams to be converted into fixed-length vectors that can be directly used with standard machine learning models.

2.7. Relation to Image Data.

In image analysis, topological structures arise from spatial arrangements of pixels and their intensity values. In particular, connected components correspond to homogeneous regions, while higher-dimensional features capture more complex geometric patterns such as boundaries and cavities.

Persistent homology provides a way to summarize these structures across multiple intensity thresholds or spatial scales. In the context of CT images, such descriptors can capture structural patterns of lung regions, including fragmented intensity regions and irregular boundaries associated with COVID-19 infection.

These concepts form the basis for the feature construction pipeline used in this paper, where images are transformed into point clouds and topological summaries are extracted for classification.

3. METHODS

Our framework follows a geometric approach in which CT images are embedded into a three-dimensional space using pixel coordinates and grayscale intensities. To improve computational efficiency while preserving structural information, we construct a Feature Point Cloud (FPC) representation. Persistent homology is then computed from both the point cloud and the image grid, and the resulting topological features are vectorized using persistence landscapes for classification.

3.1. Feature Point Cloud Construction.

Definition 3.1 (Point Cloud). *Let M be a grayscale image defined on a pixel grid $\{1, \dots, m\} \times \{1, \dots, n\}$ with intensity values $p_{ij} \in [0, 255]$. The associated point cloud is*

$$X_M = \{(i, j, p_{ij}) \in \mathbb{R}^3\}.$$

Each image is first resized to 256×256 and normalized to the range $[0, 255]$. The image is then embedded into $\mathbb{R}^3$ as

$$P = \{(x, y, I(x, y)) : (x, y) \in \Omega\}.$$

Since the raw point cloud contains one point per pixel, it is computationally expensive for persistent homology. Therefore, we construct a reduced representation using an intensity-based covering.

Definition 3.2 (Feature Point Cloud). *Let $\mathcal{U} = \{U_1, \dots, U_{15}\}$ be a collection of overlapping intervals covering $[0, 255]$ with an overlap ratio of 0.3. For each interval U_i , define*

$$\Omega_i = \{(x, y) \in \Omega : I(x, y) \in U_i\}.$$

The spatial coordinates in Ω_i are grouped using hierarchical agglomerative clustering with single linkage and Euclidean distance in the image plane, using a distance threshold $\tau = 2.0$. Clusters with fewer than three points are discarded to reduce noise.

Each cluster $C \subseteq \Omega_i$ is represented by its centroid

$$c(C) = \left(\frac{1}{|C|} \sum_{(x,y) \in C} x, \frac{1}{|C|} \sum_{(x,y) \in C} y, \frac{1}{|C|} \sum_{(x,y) \in C} I(x, y) \right).$$

To ensure computational tractability, the number of feature points is limited to 1000 using farthest point sampling, which preserves the geometric distribution of points while reducing redundancy.

The Feature Point Cloud is defined as

$$X_I = \{c(C) : C \text{ is a cluster from some } \Omega_i\}.$$

This construction reduces the number of points while preserving essential spatial and intensity-based structures of the image.

An example of the Feature Point Cloud is shown in Figure 1.

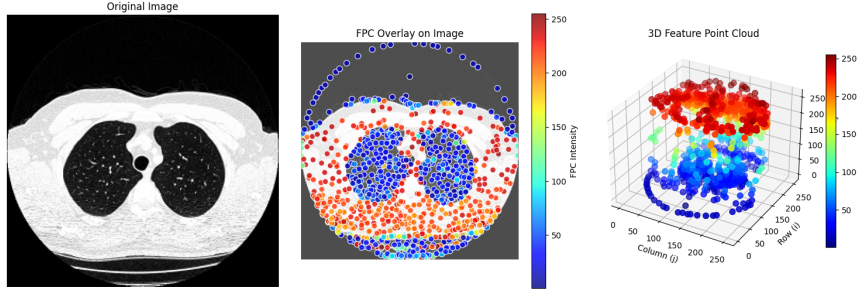


FIGURE 1. Construction of the Feature Point Cloud (FPC) from a CT scan image. Left: Original CT image. Middle: Spatial distribution of clustered feature points overlaid on the image. Right: Three-dimensional point cloud representation $(x, y, I(x, y))$ in $\mathbb{R}^3$.

3.2. Lower-Star Filtration.

In addition to the point cloud representation, we construct a filtration directly on the image grid. Each pixel is treated as a vertex, and edges are added according to 8-neighborhood adjacency, producing a cubical complex.

Let $f : K^{(0)} \rightarrow \mathbb{R}$ be defined by $f(v) = I(v)$. The lower-star filtration is given by

$$K_t = \{\sigma \in K : \max_{v \in \sigma} f(v) \leq t\}.$$

This filtration captures the evolution of connected components based on intensity thresholds in the image.

3.3. Vietoris–Rips Filtration.

Persistent homology is computed on the Feature Point Cloud using the Vietoris–Rips filtration implemented in the Giotto-TDA library. The filtration is constructed over a range of scales, allowing the extraction of multiscale topological features without fixing a single scale parameter.

Persistence diagrams are computed for homology dimensions

$$H_0, H_1, H_2.$$

Together with the diagram obtained from the lower-star filtration, we obtain four persistence diagrams per image.

To reduce the influence of noise, topological features with very small persistence are filtered out using a threshold.

Representative persistence diagrams are shown in Figure 2.

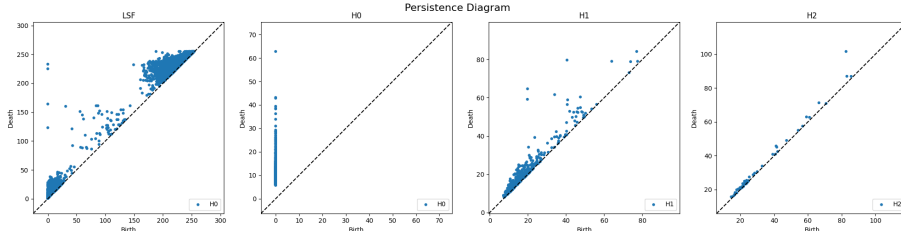


FIGURE 2. Persistence diagrams obtained from the Feature Point Cloud and lower-star filtration. Each point (b, d) represents a topological feature with birth time b and death time d . The diagrams correspond to H_0, H_1, H_2 (Vietoris–Rips filtration), and H_0 from the lower-star filtration.

3.4. Persistence Landscape Vectorization.

Each persistence diagram is transformed into a persistence landscape. We retain the first five landscape layers and sample each layer at 100 uniformly spaced points, resulting in 500 features per diagram and a total of 2000 features per image.

This vectorization produces a fixed-length Euclidean representation that can be directly used in standard machine learning models.

The corresponding persistence landscapes are illustrated in Figure 3.

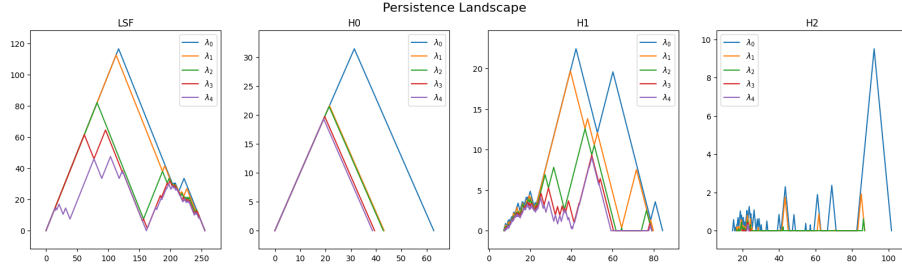


FIGURE 3. Persistence landscapes derived from the persistence diagrams. Each diagram is transformed into multiple landscape layers $\lambda_k(t)$, which are then sampled to produce a fixed-length feature vector for classification.

3.5. Implementation and Classification.

Persistent homology is computed using the Giotto-TDA library, and machine learning models are implemented using scikit-learn.

Feature vectors are standardized using statistics computed from the training data and applied consistently to validation and test sets. The dataset is split into training, validation, and testing sets with a ratio of 70:20:10 using stratified sampling.

Hyperparameters are selected using grid search with 5-fold cross-validation on the training data. For the Support Vector Machine with RBF kernel, the parameters C and γ are tuned, while for the Random Forest classifier, the number of trees, maximum depth, and splitting criteria are optimized.

The final model is retrained on the combined training and validation data and evaluated on the test set. Performance is measured using accuracy, precision, recall, F1-score, and ROC-AUC.

4. EXPERIMENTS AND RESULTS

4.1. Dataset and Experimental Setup.

The experiments are conducted on a subset of the SARS-CoV-2 CT-scan dataset [7], consisting of 1000 images with 500 COVID-19 and 500 non-COVID cases. All images are resized to 256×256 pixels prior to feature extraction.

The dataset is split into training, validation, and testing sets with a ratio of 70:20:10, respectively. The split is performed using stratified sampling to preserve class balance across all subsets, and a fixed random seed is used to ensure reproducibility.

Hyperparameters are tuned using grid search with 5-fold cross-validation on the training set. After tuning, the models are retrained on the combined training and validation data, and the final performance is evaluated on the held-out test set.

4.2. System Workflow.

The overall pipeline transforms raw CT images into topological feature representations suitable for classification:

- (1) Convert each CT image into a Feature Point Cloud.
- (2) Compute Vietoris–Rips persistence diagrams (H_0, H_1, H_2) .
- (3) Compute a lower-star persistence diagram $(H_0\text{-LSF})$.
- (4) Transform persistence diagrams into persistence landscapes.
- (5) Construct a 2000-dimensional feature vector for each image.
- (6) Perform classification using SVM and Random Forest.

This pipeline enables the transformation of spatial and intensity information into multiscale topological descriptors that can be directly integrated into standard machine learning models.

4.3. Overall Classification Performance.

Table 1 summarizes the classification performance on the test set.

TABLE 1. Overall Classification Performance on The Test Set

Metric	SVM (RBF)	Random Forest
Accuracy (%)	96.0	98.0
Precision (COVID-19)	1.00	1.00
Recall (COVID-19)	0.92	0.96
F_1 -score (COVID-19)	0.958	0.980
Precision (Non-COVID)	0.926	0.962
Recall (Non-COVID)	1.00	1.00
F_1 -score (Non-COVID)	0.962	0.980

Both classifiers achieve high performance, with Random Forest outperforming Support Vector Machine, reaching an accuracy of 98.0%. The consistently high precision for COVID-19 indicates a low false-positive rate, which is important in clinical screening scenarios.

The superior performance of Random Forest suggests that the relationship between topological features and class labels is nonlinear and involves complex feature interactions. Ensemble-based methods are better suited to capture such interactions compared to kernel-based models, which may explain the observed performance gap.

4.4. Comparison with CNN Baselines.

We compare the proposed method with several transfer-learning CNN architectures. The results are summarized in Table 2.

TABLE 2. CNN Baselines

Model	Accuracy (%)
EfficientNet-B0	90.62
EfficientNet-B2	87.65
ResNet50	91.96

The proposed TDA-based approach achieves higher accuracy than the considered CNN models. However, it is important to note that the CNN architectures are used as baseline references without extensive hyperparameter optimization. Therefore, the comparison should be interpreted as an indication of the effectiveness of topological features rather than a definitive benchmark against deep learning methods.

4.5. Ablation Study of Homological Features.

To evaluate the contribution of each homological feature, we perform an ablation study by training classifiers using each feature individually. The results are summarized in Table 3.

TABLE 3. Individual Homological Feature Performance on The Test Set

Feature	SVM Acc.	RF Acc.
H_0 (VR)	98.0%	99.0%
H_1 (VR)	76.0%	75.0%
H_2 (VR)	67.0%	76.0%
H_0 (LSF)	96.0%	95.0%

The results show that the H_0 feature derived from the Vietoris–Rips filtration provides the strongest discriminative performance, achieving 99.0% accuracy with Random Forest and 98.0% with Support Vector Machines.

This behavior can be explained by the nature of the structures captured by H_0 . In CT images of COVID-19 patients, abnormalities such as ground-glass opacities typically appear as spatially connected regions. These patterns are naturally reflected in the evolution of connected components, which are directly captured by H_0 .

In contrast, higher-dimensional features such as H_1 and H_2 , which represent loops and voids, exhibit significantly lower performance. These structures are less consistently present in lung CT images and are more sensitive to noise, sampling, and discretization effects, making them less reliable for classification.

Interestingly, combining all homological features does not improve performance compared to using H_0 alone. This suggests that higher-dimensional features may introduce redundant or noisy information, which can negatively affect the classifier.

The H_0 feature obtained from the lower-star filtration also shows strong performance, indicating that intensity-based connectivity on the image grid provides complementary structural information.

Overall, these results demonstrate that connectivity-based topological features play a dominant role in distinguishing COVID-19 CT images, while increasing topological complexity does not necessarily lead to improved classification performance.

4.6. Implementation and Reproducibility.

The implementation of the proposed method, including data preprocessing, topological feature extraction, and classification pipelines, is publicly available at:

<https://github.com/JekiSaputraaaa/covid-tda-classification>

The repository includes the full implementation as well as a Jupyter notebook that reproduces the complete experimental pipeline from raw data to final classification results.

5. CONCLUSION

In this paper, we proposed a Topological Data Analysis (TDA) framework for classifying COVID-19 CT scan images using persistent homology and persistence landscapes. The experimental results demonstrate that the proposed approach achieves strong classification performance, reaching 96.0% accuracy with Support Vector Machines and 98.0% with Random Forest on a held-out test set.

The ablation study reveals that the H_0 Vietoris–Rips feature provides the most discriminative information, outperforming both higher-dimensional features and their combination. This finding indicates that connectivity-based structures play a dominant role in distinguishing COVID-19 CT images, while higher-dimensional topological features such as loops and voids contribute limited additional information and may introduce noise.

These results highlight that increasing topological complexity does not necessarily improve classification performance, and that simpler, more stable features can be more effective in practical applications. While the proposed approach provides interpretable and competitive results, its computational cost remains higher compared to standard deep learning methods, and further evaluation on larger and more diverse datasets is required to assess its scalability and generalization.

The implementation is publicly available at <https://github.com/JekiSaputraaaa/covid-tda-classification>, supporting transparency and reproducibility of the proposed framework.

Data Availability Statement. The data used to support the findings of this study are available from the corresponding author upon request.

Declarations. The authors declare no conflict of interest.

Author Contributions. All authors have read and agreed to the published version of the manuscript. All authors contributed equally to this paper. All authors reviewed the manuscript.

Funding Statement. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Acknowledgment. The authors sincerely thank the reviewers for their valuable comments and constructive suggestions, which significantly improved the quality and clarity of this manuscript.

REFERENCES

- [1] M. Tan and Q. V. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *Proceedings of the 36th International Conference on Machine Learning (ICML)*, vol. 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114, PMLR, 2019. <https://proceedings.mlr.press/v97/tan19a.html>.
- [2] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.
- [3] G. Carlsson, “Topology and data,” *Bulletin of the American Mathematical Society*, vol. 46, no. 2, pp. 255–308, 2009. <https://doi.org/10.1090/S0273-0979-09-01249-X>.
- [4] H. Edelsbrunner, D. Letscher, and A. Zomorodian, “Topological persistence and simplification,” in *Proceedings 41st Annual Symposium on Foundations of Computer Science*, pp. 454–463, IEEE, 2000. <https://doi.org/10.1109/SFCS.2000.892133>.
- [5] S. Iqbal, H. F. Ahmed, T. Qaiser, M. I. Qureshi, and N. Rajpoot, “Classification of COVID-19 via homology of CT-Scan,” *arXiv preprint arXiv:2102.10593*, 2021. <https://arxiv.org/abs/2102.10593>.
- [6] P. Bubenik, “Statistical topological data analysis using persistence landscapes,” *The Journal of Machine Learning Research*, vol. 16, no. 1, pp. 77–102, 2015. <https://jmlr.org/papers/v16/bubenik15a.html>.
- [7] E. Soares, P. Angelov, S. Biaso, M. H. Froes, and D. K. Abe, “SARS-CoV-2 CT-scan dataset: A large dataset of real patients CT scans for SARS-CoV-2 identification,” *medRxiv*, 2020. <https://doi.org/10.1101/2020.04.24.20078584>.