World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Early detection of lung diseases using persistent homology

Md. Morshed Bin Shiraj [1, 2, *], Shanta Islam [1], Md. Hasibur Rahman [1], Md. Masum Murshed [1] and Nasima Akhter [1]

[1] *Department of Mathematics, University of Rajshahi, Rajshahi-6205, Bangladesh.*
[2] *Asian University for Women, Chattogram-4000, Bangladesh.*

World Journal of Advanced Research and Reviews, 2026, 30(03), 1914-1927

Publication history: Received on 17 May 2026; revised on 24 June 2026; accepted on 26 June 2026

Article DOI: https://doi.org/10.30574/wjarr.2026.30.3.1775

Abstract

This study investigates the integration of Persistent Homology, a fundamental tool from algebraic topology, with medical image analysis to enhance the interpretability and reliability of lung disease diagnosis from CT scans. Lung diseases remain a major global health concern and early detection is crucial for improving clinical outcomes. However, conventional diagnostic methods and many deep learning-based approaches often suffer from limitations such as subjectivity, lack of interpretability and dependency on large labelled datasets. To address these challenges, this research develops a computational framework that combines image processing techniques with Topological Data Analysis (TDA). CT scan images are pre-processed and transformed into structured data representations, where key points are selected to construct topological spaces suitable for Persistent Homology analysis. Using this approach, stable topological features such as connected components and loops are extracted and represented through persistence diagrams. Based on these topological signatures, a set of rule-based diagnostic criteria is formulated to classify lung conditions into healthy, mild and severe categories. The entire workflow is implemented within a Python-based automated framework, enabling consistent feature extraction and decision-making from raw CT images. The results demonstrate that Persistent Homology provides meaningful and interpretable descriptors of lung structure, allowing objective assessment of disease severity. This study highlights the potential of topology-based methods as a transparent and mathematically grounded alternative to traditional and deep learning approaches in medical imaging. Although the findings are promising, this work represents an initial exploration, and further validation using larger datasets and clinical studies is necessary to fully establish its diagnostic robustness. Nevertheless, the proposed framework offers a promising direction toward data-driven, interpretable, and topology-informed medical diagnostics, with potential applications in early detection and clinical decision support.

Keywords: Computed Tomography; Explainable Medical Imaging; Lung Disease; Persistent Homology; Topological Data Analysis

1. Introduction

Lung diseases-including pulmonary fibrosis, chronic obstructive pulmonary disease (COPD), pneumonia and early-stage lung cancer—continue to pose significant global health challenges, contributing substantially to morbidity and mortality worldwide [1, 7]. The effectiveness of treatment strategies is highly dependent on early detection, as delays in diagnosis often lead to rapid disease progression and poor clinical outcomes [1, 8]. At present, diagnostic practices rely predominantly on the interpretation of computed tomography (CT) scans by radiologists. While effective, this process is inherently subjective, time-intensive and dependent on the expertise of clinicians, thereby introducing variability and potential diagnostic inconsistencies [9].

In recent years, the integration of artificial intelligence, particularly deep learning, into medical imaging has demonstrated promising advancements in automated disease detection [10, 11]. However, these approaches typically

* Corresponding author: Md. Morshed Bin Shiraj

function as “black-box” models, offering limited interpretability and requiring extensive labeled datasets for training [12]. Such limitations hinder their adoption in critical clinical settings, where transparency and reliability are essential.

Within this context, Topological Data Analysis (TDA), and more specifically Persistent Homology, emerges as a mathematically rigorous and interpretable framework for analyzing complex, high-dimensional data [2, 3, 13]. Persistent Homology captures the intrinsic topological structure of data by identifying stable features such as connected components, loops, and voids across multiple scales [13, 14]. Unlike conventional image analysis techniques, these topological descriptors are inherently robust to noise and minor perturbations, making them particularly suitable for medical imaging applications [15].

The application of TDA in medical diagnostics represents a growing area of research, with notable contributions in oncology and COVID-19 imaging [16, 17]. Nevertheless, its application to the early detection of general lung diseases remains relatively underexplored, particularly in the context of developing interpretable diagnostic models. This gap underscores the need for a systematic framework that leverages topological features as meaningful biomarkers for disease classification.

In the broader landscape of medical diagnostics, the pursuit of improved accuracy, efficiency, and early detection continues to drive innovation. The convergence of advanced imaging technologies and computational methodologies has ushered in a transformative era in healthcare [5]. Among these emerging approaches, the integration of Persistent Homology with medical imaging offers a novel and insightful perspective for examining complex biological structures.

The human lung, a vital organ responsible for respiration, is highly susceptible to a wide spectrum of pathological conditions. Accurate and timely diagnosis of these conditions is critical, as it directly influences treatment effectiveness and patient prognosis [1]. Traditional diagnostic approaches, largely dependent on visual inspection, are often constrained by subjectivity and limited quantitative assessment capabilities [9].

Motivated by these challenges, this study explores the synergistic potential of Persistent Homology in lung image analysis [2, 3]. Rooted in algebraic topology, Persistent Homology provides a powerful mechanism for uncovering hidden structural patterns within complex datasets [13]. By applying this framework to lung CT images, we aim to overcome the limitations of conventional diagnostic methodologies and introduce a more objective and robust analytical paradigm.

Our approach integrates rigorous mathematical foundations with practical image processing techniques [4]. We begin by revisiting the core principles of Persistent Homology and its visualization through persistence diagrams, which serve as the theoretical backbone of our analysis [14]. Building upon this foundation, we implement a structured workflow involving dataset acquisition (e.g., as in [6]), image normalization, feature extraction and topological computation which is an extension of [18].

The outcome of this integrated methodology is the development of a Python-based computational framework capable of supporting diagnostic decision-making. This framework classifies lung conditions into categories such as healthy, mild, and severe, based on extracted topological features. By systematically quantifying and visualizing these features, the proposed approach enhances both the precision and interpretability of diagnostic outcomes.

Ultimately, this study represents a step toward bridging the gap between advanced mathematical theory and practical medical applications. By leveraging the robustness and interpretability of topological features, we contribute to the evolving landscape of data-driven healthcare, paving the way for more reliable and transparent diagnostic systems grounded in topological insights.

2. Methodology

This study adopts a systematic and computational approach to develop an interpretable framework for early-stage lung disease detection using Persistent Homology within the paradigm of Topological Data Analysis (TDA). The overall workflow consists of several interrelated stages, integrating mathematical theory, image processing and data-driven modeling.

2.1. Collection of Research Materials

The research begins with the collection of essential materials, including scientific articles, textbooks and publicly available datasets relevant to lung diseases and medical imaging. Particular emphasis is placed on acquiring high-quality

CT scan datasets (e.g., publicly available repositories such as [6]) to ensure the reliability and reproducibility of the analysis.

2.2. Literature Review and Theoretical Framework

A comprehensive literature review is conducted to examine existing methodologies in topology-based disease detection, deep learning approaches in medical imaging, and the mathematical foundations of Persistent Homology. This step establishes the theoretical basis of the study, including concepts such as simplicial complexes, filtration processes, and persistence diagrams. The review also identifies research gaps, particularly the lack of interpretable models for early-stage lung disease classification.

2.3. Model Development and Computational Environment

A computational model is developed using programming environments such as Python and R. Relevant libraries (e.g., GUDHI, Ripser, Scikit-learn and image processing libraries) are employed to implement the proposed framework. The model integrates:

- Image preprocessing techniques (noise reduction, normalization, segmentation),
- Construction of point cloud or pixel-based representations from CT images,
- Filtration methods for topological analysis,
- Computation of persistence diagrams and barcodes.

2.4. Data Acquisition and Preprocessing

CT scan data are collected and preprocessed to ensure consistency and quality. Preprocessing steps include resizing, grayscale normalization, and region-of-interest (ROI) extraction to isolate lung structures. In some cases, feature points are manually or algorithmically selected to construct suitable input representations for topological analysis.

2.5. Topological Feature Extraction

Persistent Homology is applied to the processed data to extract topological features such as connected components (0-dimensional homology), loops (1-dimensional) and higher-dimensional voids where applicable. These features are represented through persistence diagrams and transformed into numerical summaries (e.g., persistence landscapes, Betti curves or vectorized descriptors) suitable for machine learning tasks.

2.6. Model Execution and Classification

The extracted topological features are used to train and evaluate classification models. The framework categorizes lung conditions into classes such as healthy, mild, and severe. Standard machine learning algorithms (e.g., support vector machines, random forests or logistic regression) can be employed to maintain interpretability while ensuring robust performance.

2.7. Result Interpretation and Visualization

The results are analyzed both quantitatively and visually. Persistence diagrams and related topological summaries are interpreted to identify meaningful patterns associated with different lung conditions. This step emphasizes the interpretability of the model, linking topological features to anatomical and pathological structures.

2.8. Validation and Performance Analysis

To assess the robustness and reliability of the proposed persistent homology-based framework, a comprehensive validation procedure was performed using noise perturbation and threshold variation analyses. The stability of the extracted topological features was evaluated by comparing persistence diagrams generated from original and perturbed CT images using bottleneck and Wasserstein distances. Different levels of Gaussian noise were introduced to examine the sensitivity of topological representations, while threshold variation analysis was conducted to investigate the effect of segmentation parameters on feature extraction. The consistency of the obtained topological measurements was further evaluated through repeated experiments using mean values and standard deviations of the calculated distances. Finally, the computational performance of the proposed method was assessed by measuring the execution time required for preprocessing, persistent homology computation, feature extraction, and validation analysis to demonstrate its feasibility for lung CT image characterization.

2.9. Documentation and Dissemination

Finally, the findings of the study are compiled into technical reports and research manuscripts. The results are prepared for submission to peer-reviewed journals and conferences, contributing to the advancement of topology-driven medical diagnostics.

2.10. Dataset Information

The dataset utilized in this study is adopted from the work reported in [6] and is publicly available through platforms such as Kaggle and GitHub. This dataset, commonly referred to as the COVID-CT Dataset, consists of a collection of computed tomography (CT) images specifically curated to represent clinical manifestations associated with COVID-19 infection.

In total, the dataset comprises 349 CT scan images obtained from 216 patients, providing a diverse representation of infection patterns and disease severity. The images were not collected from a single clinical center; rather, they were systematically extracted from COVID-19-related scientific publications in reputable sources, including medRxiv, bioRxiv, The New England Journal of Medicine (NEJM), JAMA and The Lancet. This ensures a broad and heterogeneous sample, enhancing the generalizability of the study.

The selection process involved a careful manual review of published articles, where CT images indicative of COVID-19 were identified based on their accompanying figure captions and clinical descriptions. Only those images explicitly associated with confirmed COVID-19 cases were included in the dataset, ensuring the reliability of labels.

The dataset captures a variety of radiological patterns commonly associated with COVID-19, such as ground-glass opacities, consolidations and bilateral lung involvement. These variations make it particularly suitable for evaluating computational models aimed at detecting and classifying lung abnormalities.

For the purpose of this research, the dataset is further preprocessed to ensure uniformity in image size, intensity normalization and region-of-interest extraction. Although originally designed for COVID-19 detection, the dataset serves as a representative benchmark for testing the effectiveness of Persistent Homology in extracting topological biomarkers from lung CT images.

3. Results

The proposed methodology provides a structured and mathematically grounded framework for the analysis of CT-scanned lung images, integrating the theoretical foundations of Persistent Homology with practical image processing techniques. This hybrid approach enables the extraction of meaningful structural information from complex medical images, facilitating a more interpretable and robust assessment of lung conditions.

To ensure a solid theoretical basis, an extensive review of Persistent Homology and its graphical representation through persistence diagrams was first conducted. This step was essential in understanding how topological features—such as connected components and loops—can be used to characterize intrinsic geometric patterns within lung CT images. These mathematical constructs form the backbone of the proposed analytical pipeline and guide the subsequent computational implementation.

Following the theoretical groundwork, a dataset of CT scan images was acquired from publicly available sources, as described earlier. From this dataset, a representative image corresponding to a healthy lung was carefully selected to serve as a baseline or reference model for comparative analysis (see Fig. 1). This reference image plays a crucial role in distinguishing normal anatomical structures from pathological deviations observed in diseased lungs.

The next stage involved preprocessing and feature preparation. Each CT image was standardized to ensure consistency in resolution and intensity values. Subsequently, a semi-manual approach was adopted for feature localization, where significant regions of interest (ROIs) within the lung fields were identified. Through manual selection and point placement, key structural regions—such as boundaries of lung tissues and areas exhibiting potential abnormalities—were marked. These selected points effectively transform the image into a spatial dataset suitable for topological analysis.

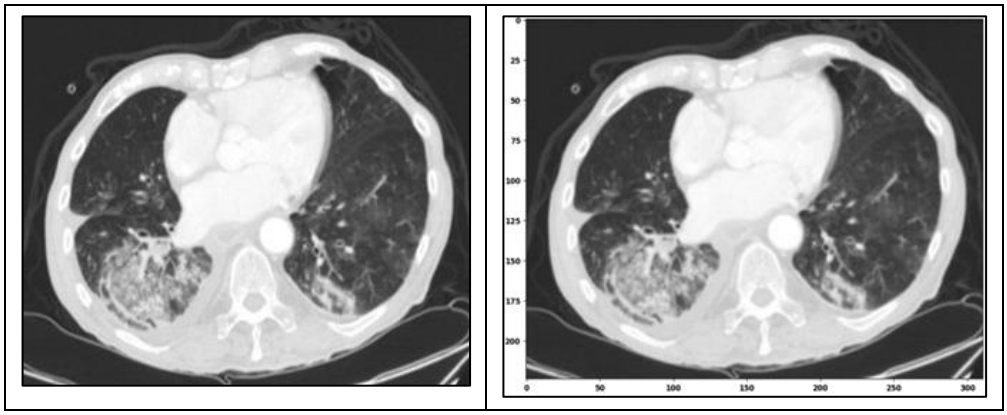

Figure 1 The selected CT scan of lung in good health (left) and the image after scaling (right)

The resulting point cloud representation of a healthy lung, illustrated in Fig. 2, serves as the input for the computation of Persistent Homology. This step is particularly important, as it bridges the gap between raw image data and abstract topological structures. By focusing on carefully chosen regions, the method ensures that the extracted features are both relevant and computationally efficient, reducing the influence of noise and irrelevant background information.

This initial phase of implementation demonstrates the feasibility of applying topological techniques to medical imaging data. The combination of a well-defined reference image, systematic preprocessing, and structured point selection establishes a reliable foundation for subsequent stages, including persistence diagram generation, feature extraction and classification. Moreover, the use of interpretable geometric representations highlights the potential of Persistent Homology to provide deeper insights into lung morphology, beyond what is typically achievable with conventional image analysis methods.

Overall, this stage lays the groundwork for a comparative and quantitative evaluation of lung conditions, setting the stage for further analysis of topological features and their role in disease classification.

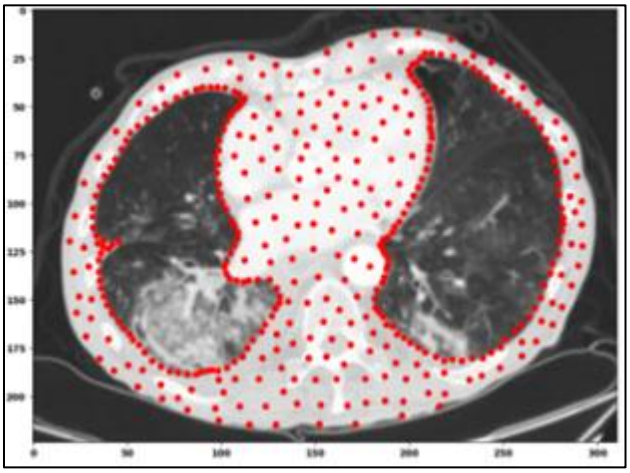

Figure 2 The red dots around the lung of the sample CT scan refers 391 points selected

Following the construction of the point cloud representation from the CT scan images, the Persistent Homology framework was applied to extract intrinsic topological features from the data. This stage constitutes the core analytical component of the proposed methodology, where geometric image information is transformed into mathematically meaningful topological descriptors.

Using the filtration process, the evolution of topological structures was tracked across multiple scales. In particular, the analysis focused on identifying key features such as connected components (0-dimensional homology) and loops (1-dimensional homology). These features provide important insights into the underlying spatial organization of lung structures, capturing both global connectivity and local structural variations.

The results of the Persistent Homology computation are visualized through a persistence diagram, as shown in Fig. 3. The diagram summarizes the birth and death of topological features across different filtration scales, allowing for the identification of stable and significant structures within the dataset.

From the persistence diagram, it is observed that the dataset exhibits one prominent connected component, indicating a dominant cohesive structure within the selected region of the lung image. This suggests that the overall lung region, after preprocessing and segmentation, forms a single unified topological space, which is consistent with anatomical expectations.

In addition, the persistence diagram reveals two significant loops, representing stable 1-dimensional topological features. These loops may correspond to structural variations or internal boundary formations within the lung region, which are not easily detectable through traditional pixel-based image analysis methods. The persistence of these loops across filtration scales indicates that they are not noise-induced artifacts but rather meaningful structural characteristics of the underlying image data.

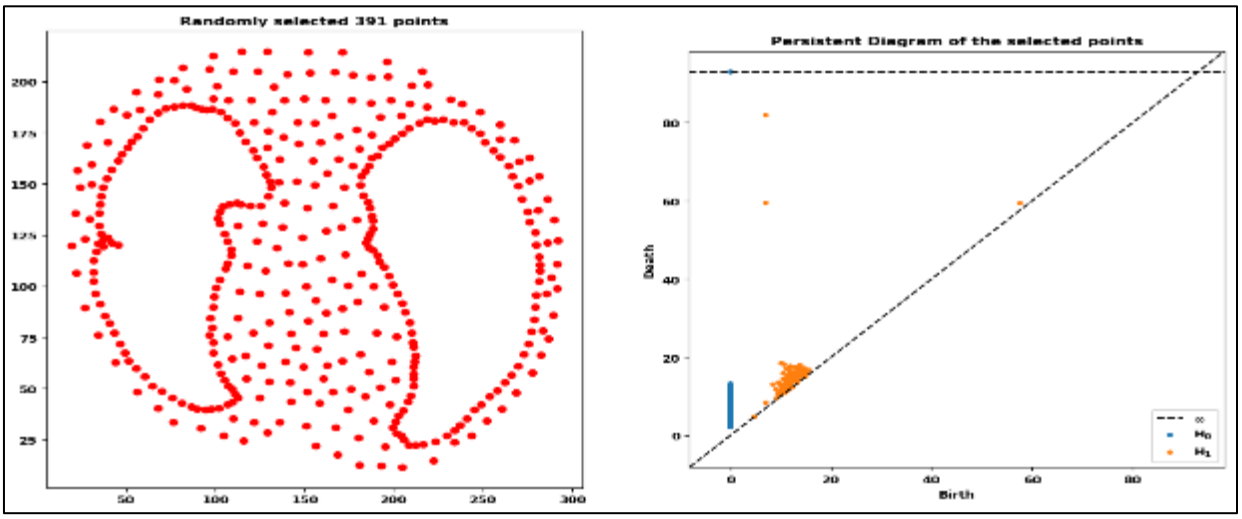

Figure 3 The output of the computation of persistent homology of the selected points (left) has been shown in the persistent diagram (right)

The identification of these topological signatures demonstrates the capability of Persistent Homology to capture multi-scale structural patterns in CT scan images. Unlike conventional feature extraction techniques that rely on intensity or texture-based information, this approach provides a shape-based representation that is robust to noise and minor perturbations.

Overall, the results presented in Fig. 3 validate the effectiveness of the proposed TDA-based framework in revealing hidden structural properties of lung CT images. The extracted topological features form a crucial basis for subsequent quantitative analysis and classification of lung conditions.

To further enhance the reproducibility and efficiency of the proposed methodology, a dedicated Python-based computational framework was developed. This framework automates the key stages of image preprocessing, feature extraction, and topological analysis, thereby reducing manual intervention and improving consistency in diagnostic interpretation.

Within this framework, each input CT scan image is first converted into a grayscale representation, which simplifies intensity-based processing while preserving essential structural information. Following this conversion, thresholding techniques are applied to define the regions of interest (ROIs) corresponding to lung structures. In this study, threshold values of 100 and 160 were employed to investigate the sensitivity of feature extraction under different intensity cutoffs, as illustrated in Fig. 4.

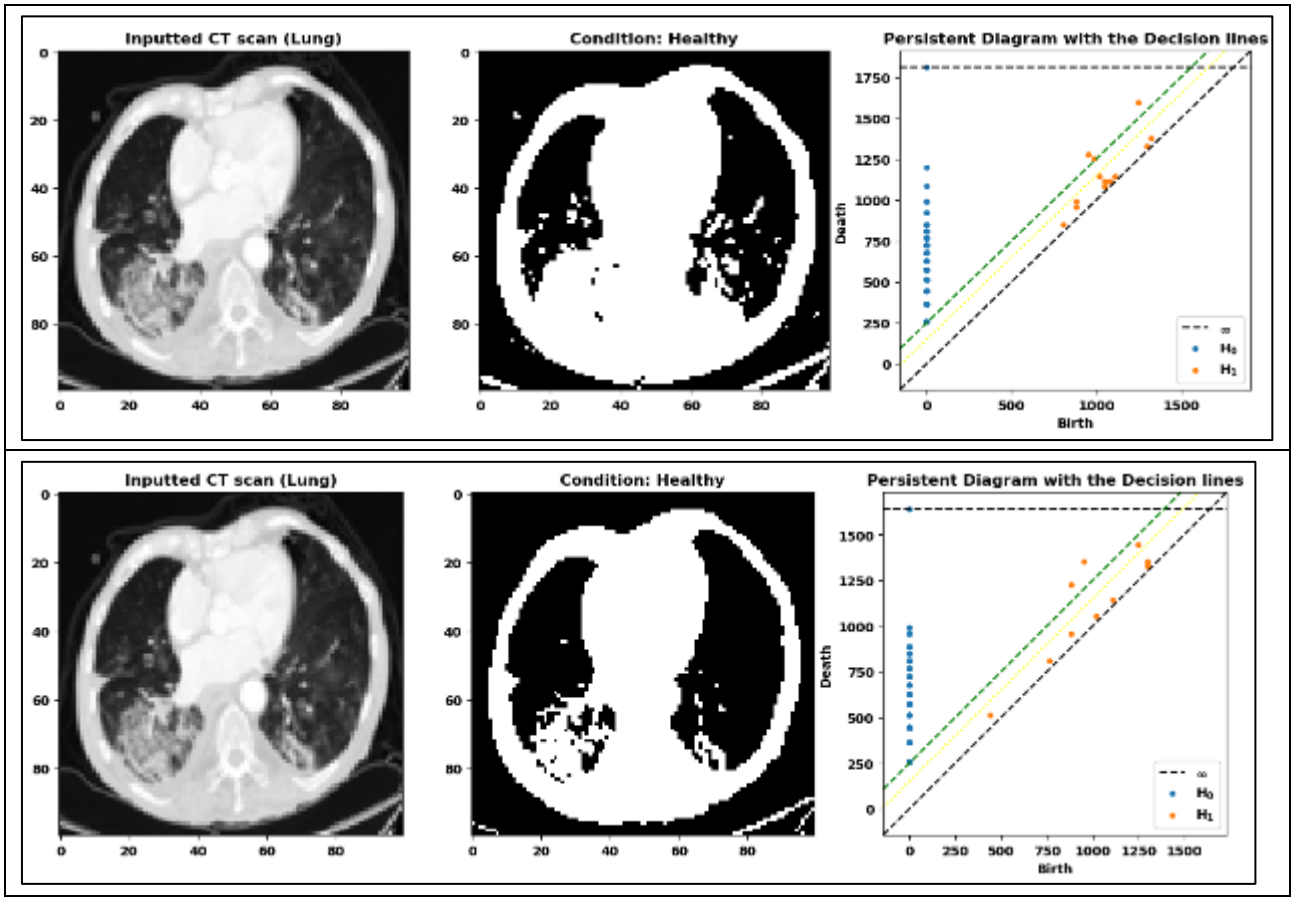

Figure 4 The proposed framework of examining the health condition of a lung have been executed over a healthy lung and has been detected as healthy by defining (a) white above 100 pixels (above) and (b) above 160 pixels (below)

Based on the identified ROIs, sample pixel points are systematically generated to construct a representative point cloud. These points serve as the foundational input for subsequent topological computations. This automated point sampling strategy ensures that the spatial distribution of lung regions is adequately captured while maintaining computational efficiency for Persistent Homology analysis.

Once the point cloud is constructed, the Persistent Homology algorithm is applied to compute topological invariants across multiple scales. The resulting persistence diagrams, also shown in Fig. 4, provide a compact representation of the evolving topological structures within the thresholded lung regions.

The analysis of these diagrams reveals the presence of three significant loops within the grayscale-transformed lung image. These loops represent stable 1-dimensional homological features that persist across a range of filtration values, indicating that they are structurally meaningful rather than artifacts of noise or thresholding variation. The emergence of multiple loops suggests the existence of complex internal structural variations within the lung region, which may correspond to subtle textural or pathological changes.

This observation further demonstrates the capability of the proposed framework to detect multi-scale structural complexity in medical images. By integrating threshold-based segmentation with topological analysis, the framework successfully bridges classical image processing techniques with modern algebraic topological methods.

Overall, the results confirm that the Python-based pipeline not only automates the diagnostic workflow but also enhances the interpretability of lung CT image analysis. The ability to consistently extract stable topological features such as loops provides a strong foundation for subsequent classification tasks and supports the potential of Persistent Homology as a reliable tool in medical image diagnostics.

Furthermore, to translate the extracted topological features into a clinically interpretable decision-making framework, a set of diagnostic rules was established based on the behavior of the 1-dimensional homology group H1, particularly

the number and persistence of loops identified in the persistence diagrams. This step is crucial for bridging the gap between abstract topological descriptors and practical medical interpretation.

In the case presented in Fig. 3, where three significant loops were detected, a rule-based classification strategy was defined using thresholded persistence values (visualized through reference boundary levels such as the green and yellow lines). Specifically, the diagnostic criteria are formulated as follows:

- A higher number of robust loops (lifespan $\geqslant$ 240) suggest a healthier lung structure. A lung is considered healthy if more than two loops (H1 features) persist above the green threshold, indicating stable and strong topological structures consistent with normal lung morphology.
- Fewer robust loops, but more subtle loops (lifespan $\geqslant$ 150), may indicate a mild condition. A lung is classified as mildly affected if at least one loop persists above the green threshold and more than three loops persist above the yellow threshold, reflecting moderate structural deviations from the healthy baseline.
- Otherwise, if these conditions are not satisfied, the lung condition is categorized as severe, indicating significant disruption of the underlying topological structure.

This rule-based interpretation provides a transparent and mathematically grounded mechanism for converting persistence diagrams into clinically meaningful diagnostic categories. Unlike black-box classification models, this approach ensures interpretability by directly linking topological persistence to disease severity.

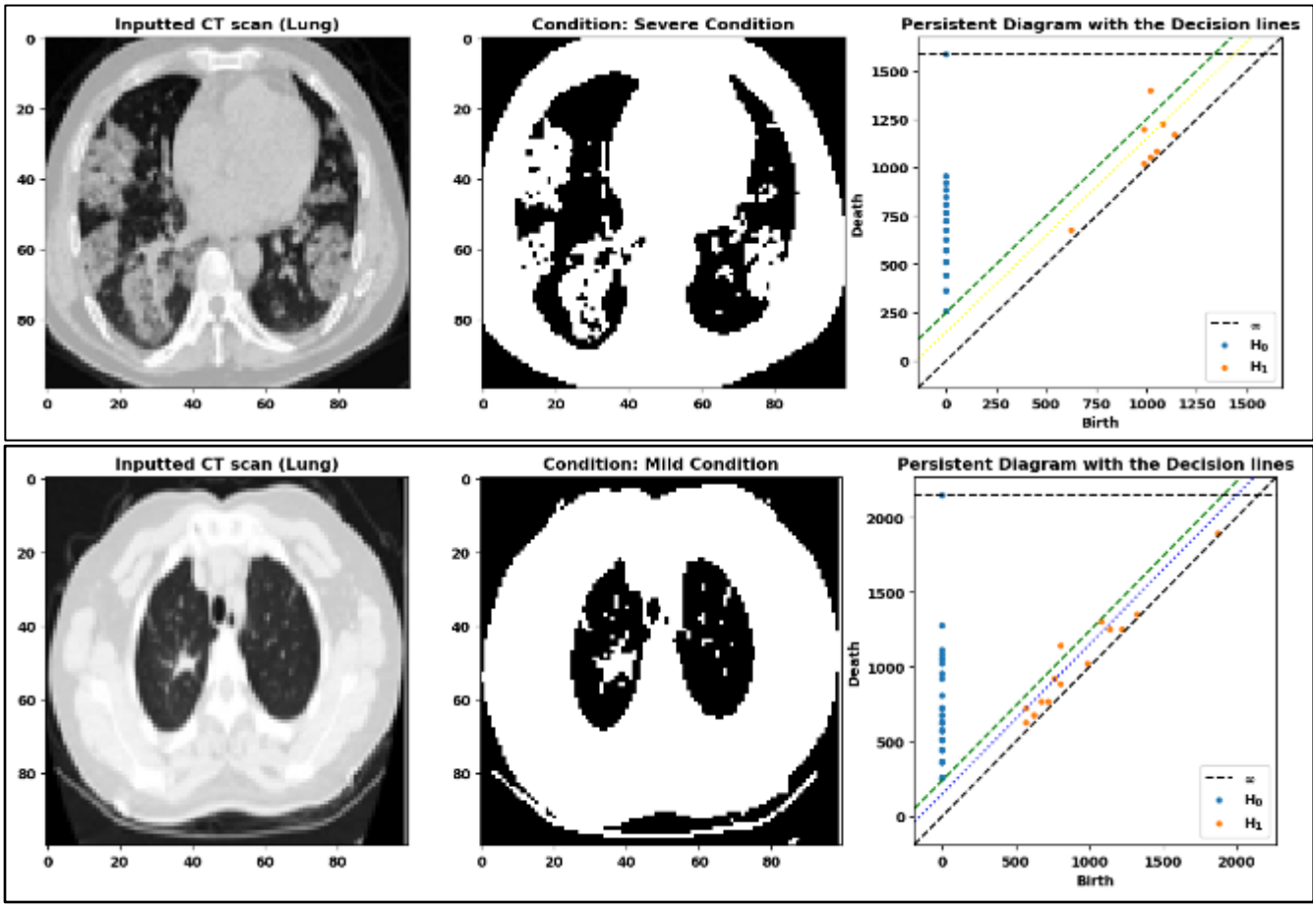

Figure 5 An example of detecting severe and mild conditioned lung from its CT scan

To further validate the robustness of the proposed framework, a similar computational procedure was applied to a COVID-19 affected lung CT scan, as illustrated in Fig. 5. In this case, white pixel intensities within the range of 140 to 255 were extracted to isolate regions of high radiological relevance, such as infection-induced opacities. The same pipeline of thresholding, point cloud generation, and Persistent Homology computation was executed to analyze the structural alterations in the infected lung tissue.

The resulting topological signatures exhibited noticeable deviations from the healthy case, reinforcing the sensitivity of the proposed method to pathological changes. These findings demonstrate that variations in loop structures and their persistence can effectively reflect underlying disease conditions, including COVID-related abnormalities.

Overall, these results provide strong evidence for the applicability of Persistent Homology-based methods in lung condition assessment. By enabling interpretable diagnostic rules grounded in topological features, the proposed framework offers a novel and promising direction for automated and explainable medical image analysis.

4. Discussion

The experimental results obtained from the proposed framework demonstrate the effectiveness of integrating Persistent Homology with CT scan image analysis for lung disease assessment. Across all examined cases, the methodology consistently succeeded in extracting stable and interpretable topological features that reflect meaningful structural variations within lung tissues.

The initial analysis of healthy lung CT images revealed a relatively simple and stable topological structure characterized by a small number of persistent components and loops. In contrast, as demonstrated in the case studies, variations in loop structures and their persistence levels were observed when thresholding parameters were adjusted and when pathological images were analyzed. These variations highlight the sensitivity of topological features in capturing subtle morphological differences that may not be easily detectable using traditional pixel-based or intensity-based image analysis techniques.

A key contribution of this study lies in the transformation of persistence diagram outputs into a rule-based diagnostic framework. By defining thresholds on the persistence of H1 features (loops), the study introduces an interpretable decision-making mechanism. The classification rules—distinguishing healthy, mild and severe conditions based on the number and stability of loops—provide a transparent mapping from abstract mathematical structures to clinically meaningful categories. This interpretability is a significant advantage over conventional deep learning approaches, which often lack transparency in their decision processes.

The application of the framework to COVID-19 affected lung CT scans further validates its robustness. In these cases, distinct topological patterns were observed when compared to healthy lung images, particularly in terms of loop formation and persistence distribution. The use of intensity-based thresholding (e.g., 140–255 for white pixel extraction) enabled the isolation of infection-related regions, which contributed to the detection of altered topological signatures. These findings suggest that Persistent Homology is capable of capturing disease-induced structural distortions in lung tissue representation.

Moreover, the integration of a Python-based computational pipeline ensured reproducibility and scalability of the proposed approach. The automation of preprocessing, thresholding, point cloud generation, and topological computation significantly reduces manual intervention and enhances consistency in analysis. This makes the framework suitable for further extension into large-scale clinical datasets.

From a comparative perspective, the proposed method offers several advantages over conventional approaches. Unlike deep learning models, which require large labeled datasets and operate as black-box systems, the TDA-based framework provides mathematically interpretable features, is robust to noise, and performs effectively even with limited data availability. However, it should be noted that the method still depends on careful selection of parameters such as threshold values and filtration settings, which may influence the resulting topological structures.

Thus, the results strongly indicate that Persistent Homology can serve as a powerful tool for lung image analysis, enabling both accurate and interpretable diagnostic insights. The combination of mathematical topology with computational image processing opens a promising direction for future research in explainable medical AI. Further improvements may include fully automated ROI selection, integration with larger multi-class datasets, and hybridization with machine learning models to enhance classification performance while preserving interpretability.

4.1. Stability Validation of Persistent Homology Features:

To evaluate the robustness of the extracted topological features, stability analysis was performed under two types of perturbations: (i) controlled noise variation and (ii) threshold variation during image preprocessing for the CT scan data shown in Fig. 5 as a sample case. Since persistent homology is theoretically stable under small perturbations, this

analysis investigates whether the persistence-based features extracted from lung CT images maintain consistent topological characteristics under realistic imaging variations.

4.1.1. Noise-Based Stability Analysis

Gaussian noise was introduced into the preprocessed CT images with different noise levels ($\sigma = 1, 2, 5, 10, 15, 20$). For each noise level, the persistence diagrams were recomputed and compared with the original persistence diagram using bottleneck and Wasserstein distances.

The bottleneck distance between two persistence diagrams is defined as $d_B = d(D_1, D_2)$, where D_1 represents the original persistence diagram and D_2 represents the persistence diagram obtained from the perturbed image. It measures the maximum displacement required to match corresponding topological features between the two diagrams.

The obtained bottleneck distances are presented in Table 1.

Table 1 Mean Bottleneck Distance for different Gaussian Noises

Noise level (σ)	Mean Bottleneck Distance
1	6.27
2	11.98
5	28.21
10	48.77
15	66.09
20	86.47

The results demonstrate a gradual increase in bottleneck distance with increasing noise intensity. Specifically, the distance increased from 6.27 at (σ =1) to 86.47 at (σ =20). This monotonic behavior indicates that the topological structure changes progressively as the image perturbation becomes stronger rather than undergoing abrupt instability.

These observations suggest that the extracted persistence features remain relatively stable under low and moderate noise conditions. The gradual degradation of the topological similarity agrees with the stability property of persistent homology, where small perturbations in the input space led to bounded variations in the corresponding persistence diagrams.

Table 2 Coefficient of Variation for different Gaussian Noises

Noise level (σ)	Coefficient of Variation
1	12.8%
2	9.0%
5	8.5%
10	11.7%
15	11.9%
20	7.5%

To further evaluate the reliability of these measurements, the standard deviation of bottleneck distances was analyzed. For example, at (σ =10), the mean bottleneck distance was 48.77 with a standard deviation of 5.71, giving a coefficient of variation:

$$CV = \frac{5.71}{48.77} \approx 11.7\%.$$

The calculated coefficients of variation for all noise levels are tabulated in Table 2.

The relatively small variation values indicate that the observed changes are mainly due to controlled noise perturbations rather than random fluctuations, confirming the reproducibility of the stability measurements.

4.1.2. Wasserstein Distance Analysis:

In addition to bottleneck distance, Wasserstein distance was calculated to measure the overall displacement of all persistence points between the original and perturbed diagrams. Unlike bottleneck distance, which considers only the largest feature displacement, Wasserstein distance evaluates the cumulative changes across all topological features.

The obtained Wasserstein distances have been shown in Table 3.

Table 3 Mean Wasserstein Distance for different Gaussian Noises

Noise level (σ)	Mean Wasserstein Distance
1	88.29
2	174.77
5	394.79
10	655.91
15	848.79
20	952.04

The Wasserstein distance shows a consistent increasing trend as noise intensity increases. The values rise from 88.29 at $\sigma = 1$ to 952.04 at $\sigma = 1$, indicating that stronger perturbations gradually modify the global topological organization of the lung CT images.

The larger magnitude of Wasserstein distances compared with bottleneck distances is expected because Wasserstein distance considers the collective movement of all persistence features, whereas bottleneck distance only considers the most significant change. Therefore, these two complementary measurements demonstrate that the proposed persistent homology framework captures stable yet sensitive topological characteristics from CT images.

4.1.3. Threshold Stability Analysis:

The influence of segmentation threshold selection on the extracted topological features was evaluated by varying the grayscale threshold values. The persistence diagram obtained using threshold value 160 was considered as the reference diagram, and the bottleneck distance between this reference and diagrams generated using other thresholds was calculated.

The obtained results are mapped in Table 4.

As expected, the reference threshold (160) produced a bottleneck distance of zero because it generates the same persistence diagram used for comparison. The neighboring threshold values of 140 and 180 resulted in relatively smaller distances (94.26 and 89.00, respectively), indicating that the extracted topological features remain relatively consistent within this threshold range.

However, larger deviations from the reference threshold produced greater changes in topology. For example, thresholds 80, 100, 120, and 200 resulted in distances above 160. This behavior is expected because threshold selection directly influences the binary representation of the lung structure by modifying connected components, cavities, and boundaries, which are the fundamental structures captured by persistent homology.

The identical bottleneck values observed for thresholds 100, 120, and 200 suggest that either the resulting binary images contain similar topological structures or that the dominant persistence features governing the bottleneck matching remain unchanged. Further analysis of pixel intensity distributions and segmented image structures can be used to investigate this behavior.

Table 4 Bottleneck Distance for different Threshold Values with reference threshold at 160

Threshold Value	Bottleneck Distance
80	164.61
100	189.86
120	189.86
140	94.26
160	0.00
180	89.00
200	189.86

Therefore, stability experiments demonstrated that the persistence diagrams varied smoothly with increasing Gaussian noise. The mean bottleneck distance increased from 6.27 at noise level σ =1 to 86.47 at σ =20, while the Wasserstein distance increased from 88.29 to 952.04. These results indicate that the extracted topological features are robust to moderate perturbations and degrade progressively under severe noise. Threshold sensitivity analysis further revealed local stability around the selected segmentation threshold (160), with bottleneck distances below 100 for thresholds 140 and 180, whereas larger deviations from the reference threshold produced substantially greater topological changes.

5. Conclusion

This study demonstrates that Persistent Homology, within the framework of Topological Data Analysis (TDA), provides a powerful, interpretable, and mathematically grounded approach for the early detection of lung diseases from CT scan images. By systematically extracting stable topological features such as connected components and loops, the proposed methodology successfully transforms complex medical imaging data into meaningful and quantifiable descriptors that can be directly linked to clinical interpretation.

Through the integration of image processing techniques and Persistent Homology, a complete computational pipeline was developed from preprocessing and region selection to filtration, persistence diagram construction, and feature-based diagnostic rule formulation. The resulting framework enables the conversion of abstract topological signatures into explicit diagnostic decisions, categorizing lung conditions into healthy, mild, and severe classes. This rule-based interpretability represents a significant advantage over conventional deep learning models, which often lack transparency in decision-making.

The implementation of a Python-based automated framework further enhances the practicality and reproducibility of the approach. It reduces manual intervention in image processing, ensures consistency in feature extraction, and supports efficient computation of topological descriptors. The experimental observations confirm that variations in loop structures and their persistence values are strongly associated with differences in lung conditions, demonstrating the diagnostic relevance of topological features.

Importantly, this research highlights the potential of Persistent Homology to address key limitations in existing medical imaging approaches, particularly the dependence on large labeled datasets and the lack of interpretability in black-box AI models. By focusing on mathematically interpretable topological biomarkers, the proposed method enhances diagnostic transparency, reduces subjectivity in radiological interpretation, and supports more reliable clinical decision-making.

Although the results are promising, this study should be considered an initial step toward a broader application of TDA in medical diagnostics. The current framework relies on threshold selection and semi-manual preprocessing, which may introduce variability. Therefore, further research involving larger and more diverse datasets, fully automated segmentation techniques, and clinical validation is necessary to strengthen the robustness and generalizability of the model.

In conclusion, the findings suggest that Persistent Homology has significant potential to transform lung disease diagnosis by providing a robust, interpretable, and computationally efficient alternative to traditional and deep

learning-based methods. Its ability to extract stable topological structures from CT scans offers a new perspective in medical image analysis, paving the way for more transparent, accessible, and reliable diagnostic systems.

Compliance with ethical standards

Acknowledgments

The authors gratefully acknowledge the referees for their valuable comments and constructive suggestions, which have significantly improved this manuscript. This work was supported by a research grant from the Dean, Faculty of Science, University of Rajshahi (Grant No. A-155/5/52/RU. /Science-2/2025–2026). The authors gratefully acknowledge this support.

Disclosure of conflict of interest

The authors declare that there is no conflict of interest.

Statement of ethical approval

This computational proof-of-concept study used de-identified, publicly available CT images and did not recruit new participants. No identifiable patient information was accessed by the authors.

Data availability

The COVID-CT dataset used as source material is publicly available as described by Yang et al. [6].

Funding

This work was financial supported by the Dean of the Faculty of Science, University of Rajshahi, under the research "Grant No. A-155/5/52/RU. /Science-2/2025-2026".

References

[1] Agustí, Àlvar, Bartolomé R. Celli, Gerard J. Criner, David Halpin, Antonio Anzueto, Peter J. Barnes, Jean Bourbeau, et al. (2023). Global Initiative for Chronic Obstructive Lung Disease 2023 Report: GOLD Executive Summary. Archivos De Bronconeumología 59, no. 4 (April): 232–248. https://doi.org/10.1016/j.arbres.2023.02.009.

[2] Bukkuri, A., Andor, N., and Darcy, I. K. (2021). Applications of Topological Data Analysis in Oncology. Frontiers in Artificial Intelligence, 4:659037. https://doi.org/10.3389/frai.2021.659037.

[3] Moon, C., Li, Q., and Xiao, G. (2023). Using persistent homology topological features to characterize medical images: Case studies on lung and brain cancers. The Annals of Applied Statistics, 17(3): 2192–2211.

[4] González, R. C., Woods, R. E., and Masters, B. R. (2009). Digital Image Processing (3rd ed.). Journal of Biomedical Optics, 14:029901.

[5] Waheed, S., Tahir, M. J., Ullah, I., Alwalid, O., Irshad, S. G., Asghar, M. S., and Yousaf, Z. (2022). The impact of dependence on advanced imaging techniques on the current radiology practice. Annals of Medicine and Surgery, 78:103708. https://doi.org/10.1016/j.amsu.2022.103708.

[6] Yang, X., He, X., Zhao, J., Zhang, Y., Zhang, S., and Xie, P. (2020). COVID-CT-dataset: a CT scan dataset about COVID-19. arXiv preprint arXiv:2003.13865.

[7] Zhai, Y., Zhu, C., Zhu, T., Song, W., Tang, Y., Jiang, L., Ruan, F., Xu, Z., Li, L., Fu, X., Liu, D., Chen, A., and Wu, Q. (2025). Global, regional, and national burden of chronic respiratory diseases, 1990–2021 and predictions to 2035: Analysis of data from the Global Burden of Disease Study 2021. Annals of Medicine, 57(1):2530225. https://doi.org/10.1080/07853890.2025.2530225.

[8] Siegel, R. L., Miller, K. D., Wagle, N. S., and Jemal, A. (2023). Cancer statistics, 2023. CA: A Cancer Journal for Clinicians, 73(1): 17–48. https://doi.org/10.3322/caac.21763.

[9] Brady, A. P. (2017). Error and discrepancy in radiology: Inevitable or avoidable? Insights into Imaging, 8(1): 171–182. https://doi.org/10.1007/s13244-016-0534-1.

[10] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A. W. M., van Ginneken, B., and Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42: 60–88. https://doi.org/10.1016/j.media.2017.07.005.

[11] Shen, D., Wu, G., and Suk, H. I. (2017). Deep learning in medical image analysis. Annual Review of Biomedical Engineering, 19: 221–248. https://doi.org/10.1146/annurev-bioeng-071516-044442.

[12] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1: 206–215. https://doi.org/10.1038/s42256-019-0048-x.

[13] Chazal, F., De Silva, V., Glisse, M., and Oudot, S. (2016). The Structure and Stability of Persistence Modules. Springer.

[14] Otter, N., Porter, M. A., Tillmann, U., Grindrod, P., and Harrington, H. A. (2017). A roadmap for the computation of persistent homology. EPJ Data Science, 6(1), 17.

[15] Bendich, P., Marron, J. S., Miller, E., Pieloch, A., and Skwerer, S. (2016). Persistent homology analysis of brain artery trees. Annals of Applied Statistics, 10(1): 198–218. https://doi.org/10.1214/15-AOAS886.

[16] Crawford, L., Monod, A., Chen, A. X., Mukherjee, S., and Rabadán, R. (2020). Predicting clinical outcomes in glioblastoma: An application of topological and functional data analysis. Journal of the American Statistical Association, 115(531): 1139–1150. https://doi.org/10.1080/01621459.2019.1671198.

[17] Iqbal, S., Ahmed, H. F., Qaiser, T., Qureshi, M. I., and Rajpoot, N. (2025). Classification of COVID-19 via homology of CT-scan. Computers in Biology and Medicine, 193: 110226. https://doi.org/10.1016/j.compbiomed.2025.110226.

[18] Shiraj, M. M. B., Murshed, M. M., Akhter, N. (2023). Examining Lung Health through Topological Features: A CT Scan Analysis. 23rd International Mathematics Conference, May 16-17, 2024 At: Department of Mathematics, University of Rajshahi. https://www.researchgate.net/publication/380819056_Examining_Lung_Health_through_Topological_Features_A_CT_Scan_Analysis.