BMC Biomedical Engineering

RESEARCH

Open Access

# Topological data analysis and machine learning for COVID-19 detection in CT scan lung images

Rabih Assaf[1*], Abbas Rammal[2], Alban Goupil[3], Mohammad Kacim[1] and Valeriu Vrabie[3]

**Abstract**

COVID-19 has claimed the lives of thousands over the past years. Although pathogenic laboratory testing is the established standard, it carries a significant drawback with a notable rate of false negatives. Consequently, there is an urgent need for alternative diagnostic approaches to combat this threat. In response to this pressing need for accurate and parameter-free methods for COVID-19 identification, particularly within lung images, we introduce a novel approach that combines the principles of topological data analysis with the capabilities of machine learning. Our proposed methodology entails the extraction of persistent homology features from lung images, effectively capturing the intrinsic topological properties inherent in the data. These extracted persistent homology features then serve as inputs for various machine learning methods employed for classification purposes. Our primary objective is to achieve exceptional accuracy in the detection of COVID-19 all while showcasing the effectiveness of these topological features. The experimental results demonstrate that the Random Forest Classifier and the Support Vector Machine models outperform the rest, showcasing their effectiveness in classifying CT scan lung images with remarkable precision—an accuracy rate of 97.5% for the Random Forest model and an AUC score that surpasses 0.99 for the SVM. Results of the model on the same data after exclusion of the topological features and on other data with application of the same model with topological features showed the efficiency of these features in the classification task.

**Keywords** Topological data analysis, COVID-19 detection, Machine learning, Lung images

## Introduction

COVID-19, also known as the coronavirus disease 2019, is a highly contagious respiratory illness caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The outbreak of this novel coronavirus was first identified in Wuhan, China, in December 2019, and it rapidly evolved into a global pandemic. COVID-19 has since had a profound impact on public health, economies, and daily life around the world. One of the key challenges posed by COVID-19 is its rapid transmission, which has led to an urgent need for efficient and accurate methods of detection. Timely and precise identification of COVID-19 cases is crucial for controlling its spread,

*Correspondence:
Rabih Assaf
rabihassaf@usek.edu.lb
[1]Faculty of Arts and Sciences, Department of Mathematics, Holy Spirit University of Kaslik, Jounieh, Lebanon
[2]Faculty of Arts and Sciences, Mathematics and Computer Sciences Department, Lebanese American University, Beirut, Lebanon
[3]Université de Reims Champagne Ardenne, CReSTIC EA 3804, Reims 51097, France

facilitating appropriate medical care, and preventing healthcare systems from becoming overwhelmed [1, 2].

Various methods have been developed to detect COVID-19 in individuals. These methods typically fall into two main categories: diagnostic tests that directly detect the presence of the virus and tests that identify the body's immune response to the virus.

1. Molecular Tests (Virus Detection): The most common method for diagnosing COVID-19 is through molecular tests, such as the polymerase chain reaction (PCR) test. This test detects the genetic material of the virus in a patient's respiratory sample. It is highly accurate but may require specialized equipment and time for processing [3, 4].
2. Antigen Tests: Antigen tests detect specific proteins on the surface of the virus. They are quicker and less expensive than PCR tests, making them suitable for rapid screening. However, they might be less sensitive, resulting in the need for confirmation with a molecular test [5, 6].
3. Serological Tests (Antibody Detection): Serological tests detect antibodies that the immune system produces in response to the virus. These tests are useful for identifying past infections and studying the prevalence of the virus within populations. They are not suitable for early diagnosis, as it takes time for the body to produce detectable antibodies [7].
4. Chest Imaging Techniques: Imaging methods, such as X-rays and computed tomography (CT) scans of the chest, can also aid in the diagnosis of COVID-19. These techniques can reveal characteristic patterns of lung damage caused by the virus. However, they are not definitive for diagnosis and are often used in conjunction with other tests [8, 9].

In recent times, advanced technologies like artificial intelligence and machine learning have been explored to enhance the accuracy and speed of COVID-19 detection. These technologies can analyze patterns in medical images, such as lung X-rays and CT scans, to aid in early identification and triaging of COVID-19 cases [10–12]. Efforts continue to evolve in the field of COVID-19 detection, aiming to improve the accessibility, speed, and reliability of testing methods to effectively manage and mitigate the impact of the pandemic on a global scale. In the realm of machine learning, numerous recent studies and articles have observed a growing convergence between the significance of the detection of COVID-19 and the application of machine learning methods. The predominant advantage of machine learning lies in its ability to swiftly process and leverage vast volumes of data, surpassing the pace at which traditional statistical analyses can be conducted by humans.

A previous study has shown a comparative analysis of machine learning and soft computing models to predict the occurrence of COVID-19 as another method to the Susceptible-Infectious-Removed (SIR) and Susceptible-Exposed-Infectious-Removed (SEIR) models [13]. Many machine-learning models have been studied, but two have shown brilliant results (i.e., Multilayer Perceptron, MLP, and Adaptive Network-Based Fuzzy Inference System, ANFIS) [14]. Machine learning and deep learning are alternatives to humans with accurate diagnostics [15]. You can train your machine learning model using x-rays and computer tomography (CT) scans. Wang and Wong have developed a deep convolutional neural network (COVID-Net) that can diagnose COVID-19 from chest x-rays [16]. Moreover, recent studies show the potential of Artificial Intelligence and Machine Learning tools by suggesting a new model that comes with rapid and valid method SARS-CoV-2 diagnosis using Deep Convolutional Network [17, 18]. In this regard, the remarkable performance suggests the use of the convolutional neural network (Resnet-101) as an adjuvant tool for increase the accuracy of COVID-19 diagnosis. Recent studies used supervised machine learning techniques for classifying the text into four different categories COVID, SARS, ARDS and Both (COVID, ARDS). Logistic regression and Multinomial Naïve Bayes showed better results than other ML algorithms for detecting COVID-19 using clinical text data set [19].

In a comparative analysis framework between various supervised machine learning algorithms in diagnosing COVID-19 infections, Pijush Dutta implemented bagging algorithm, k-nearest neighbor, and random forest for classifying the datasets of COVID-19 [20]. Random Forest gave better results with employing accuracy of 85.71%. Haochen Yao and Nan Zhang built a COVID-19 severeness detection model based on supervised machine learning algorithms [21], and obtained the best model using the Support Vector Machine algorithm (SVM) with 28 features and overall accuracy of 81.48%. Mahbubunnabi Tamal trained a supervised machine learning algorithms to distinguish between COVID-19 and other diseases, where he found that SVM and Ensemble Bagging Model Trees (EBM) when trained on 71 radiomics features can distinguish between COVID-19 and other diseases with an overall sensitivity of 99.6% and 87.8% and specificity of 85% and 97% respectively [22]. Davide Brinati developed two machine learning classification models using hematochemical values from routine blood exams to discriminate between patients who are either positive or negative to the SARS-CoV-2 [23]: their accuracy ranges between 82% and 86%, and sensitivity between 92% and 95%.

Topological data analysis (TDA) is a methodology designed to capture the underlying geometric structure,

encompassing coarse-scale, global, and nonlinear geometric attributes, within high-dimensional datasets. Rooted in concepts from algebraic topology, TDA has exhibited remarkable success in diverse data domains, spanning from molecular to population-level information [24, 25]. TDA finds an optimal fit in the context of quantifying the architectural characteristics of prostate cancer from a prostatic perspective. In the realm of histological classification, the Gleason grading system relies on recognizing structural motifs formed by clusters of cancer cells and surrounding stroma. Even in complex higher-dimensional spaces, TDA empowers an intuitive comprehension of data patterns and their inherent architecture. Notably, the concept of persistent homology (PH) has emerged as a novel multiscale representation of topological attributes. Among the array of algebraic topology tools, persistent homology stands out as one of the most potent and computationally viable techniques for quantifying the topological traits of functions.

Our research investigates the presence of COVID-19, focusing on lung images obtained through CT scan since this kind of format is more suitable for processing computationally while maintaining sufficient resolution. To extract meaningful insights, we compute persistent homology features from these images. Leveraging machine learning techniques on these homological persistence features holds significant potential in accurately identifying the appropriate classification for COVID-19 within such images. Consequently, our objective is to enhance pathologists' comprehension of the detection of COVID-19' system by amalgamating topological data analysis with machine learning algorithms. The efficacy of our method lies in its capacity to distinguish between COVID-19 patterns from non-COVID-19 patterns with an accuracy that passes 97%. To date, there hasn't been any research that utilizes the resilience of persistent homology features in machine learning classification techniques for identifying COVID-19. Our article employs a more comprehensive array of machine learning methods, thereby expanding the horizon for more advanced outcomes. Our methodology involves the categorization of lung images into two classes. The process commences with images preprocessing. We start by image resize and normalization, enhancement then we perform lung segmentation. Subsequently, topological features are computed over the image. Machine learning techniques are then applied to predict the classes of each image.

Over time, there have been numerous proposals discussing the significance of persistent homology and others exploring the evolution of machine learning within the medical field. Yet, no one tutored the integration of these two on COVID-19 detection. Meanwhile, many methods have emerged using deep learning or machine learning to classify CT scan images as either COVID-19-affected or not. Our proposed model stands out, demonstrating exceptional performance across all aspects when compared to existing models trained on these types of images. In Table 1, we present a selection of methods applied to CT scan images along with their respective results. It is evident that achieving accuracy through the classification of topological features in the image, rather than relying solely on statistical or direct pixel values, holds great promise and underscores the potential of our approach.

## Mathematical methods

### Topological data analysis

TDA (Topological Data Analysis) extracts a significant set of topological attributes from high-dimensional datasets, which complement geometric and statistical characteristics. This presents a distinct viewpoint for machine

**Table 1** Review on methods and quantitative results for the classification of COVID-19 CT-Scan Images

| Paper | Pre-processing applied | Model applied | Performance |
|---|---|---|---|
| T. D. Pham [26] | Augmentation techniques involving random reflection, translation, and scaling | Pre-trained Convolutional Neural Network (CNN) model with weights from the ImageNet dataset. | The highest performance obtained by DenseNet201. Accuracy: 96.20%, AUC = 0.98 ± 0.03. |
| V. Shah et al [27] | Unprocessed | Pre-trained CNN model with ImageNet weight on CT scan images | The highest performance obtained by VGG19. Accuracy: 94.50%. |
| Y. Pathak et al. [28] | Unprocessed | Pre-trained CNN model with ImageNet weight | The highest performance obtained by ResNet32. Accuracy: 93.01% |
| R. Tiwari. et al [29] | Data augmentation, resizing images and channels slicing and stacking. | Channel based overlapping CNN tower architecture | The obtained accuracy is 99.40%, AUC = 0.99. |
| A. Sinha et al. [30] | Manual assessment to compromise segmentation accuracy | ML prediction model based on clinical parameters and automated CT scan features | The highest performance obtained by Random Forest Model ALLR. AUC: 0.91. |
| M. Subramanian, et al [31] | Image augmentation techniques based on "in-place" and "on-the-fly" methods | Learning without forgetting by leveraging transfer learning using CNN | The highest performance obtained by Wide ResNet. Accuracy: 98.12% |
| E.H. Lee, et al [32] | Scaling, data augmentation, applying a clipping function that truncates all Hounsfield unit intensity values above a fixed pre-determined value. | CNN that uses the entire chest CT volume applied on data in different countries | Highest accuracy: 93.2%, Highest AUC: 0.994 |


**Fig. 1** The flowchart for the computation of persistent homology classes

learning. Conversely, features derived from conventional topological models like cell complexes retain the overall inherent structural information. However, they often overly simplify the structure, and their use in quantitative characterization is limited. In a specific instance, in [33], the authors discuss the concept of Computational Topology, which integrates topological tools into established algorithms for data manipulation. In this study, our focus lies in employing persistent homology, a well-researched and widely applied tool in TDA.

### Persistent homology

Persistent Homology (PH) stands as one of the extensively explored and utilized tools within Topological Data Analysis (TDA). It possesses the ability to incorporate geometric insights into topological characteristics. This allows for the tracking of topological measurements concerning the "birth" and "death" of distinct elements like isolated components, circles, rings, loops, pockets, voids, or cavities across all geometric scales. The methods elucidated by the authors in [34, 35] explain deeply how persistent homology is computed through nested topological spaces. The method introduced here adheres to a structured procedure, commencing with the conversion of image pixels into a cubical complex, denoted as the "pixels complex." Subsequently, the filtration scheme is devised to construct an ordered sequence of these pixels' complexes. At this point, persistent homology computations are applied to this filtration scheme, allowing for the identification of homology classes that exhibit the most prolonged persistence throughout the filtration process. In Fig. 1, a flowchart shows the steps of computation of persistent homology, starting from the input image to obtaining the persistent homology classes of dimensions 0 and 1.

#### *Topological transformation to pixels' complex*

Cubical complexes provide a structured space for analyzing 2D grayscale images. In this framework, we create complex structures using cubes, squares, edges, and vertices. Specifically, when dealing with grayscale images, we represent individual pixels as vertices (0-cells) in the complex [36, 37]. The relationships between pixels are captured by unit length edges (1-cells), connecting pixels with coordinates that differ by one unit along a single axis. Moreover, unit squares (2-cells) are constructed by connecting its four-unit edges. This construction process

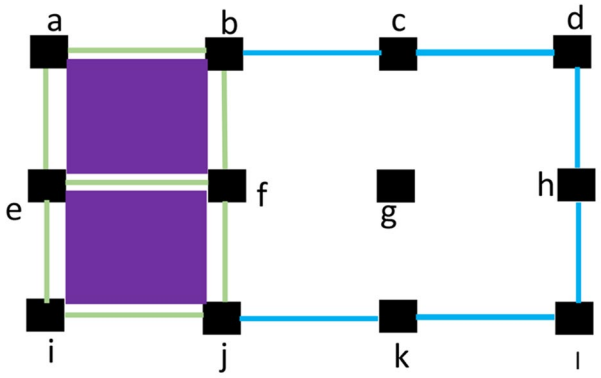


**Fig. 2** An example of a pixels' complex composed from vertices, edges and squares

results in a visual representation as in Fig. 2, where black-colored vertices represent pixels. These vertices can be connected by edges (1-cells). Squares (2-cells) are created using the 4 surrounding edges. The pixel's complex in Fig. 2 represents an example of an instance of the topological space that will be created during the filtration scheme described later in section "Filtration of the pixels' complex". It represents a 3*4 image, with 12 pixels-vertices represented by letters from a to l. Edges can connect each two vertices vertically or horizontally, (ab) or (hl) for example. Squares are formed by the 4 surrounding edges as the square (abfe) is formed by (ab), (bf), (fe) and (ea). This approach allows us to create a comprehensive complex that encapsulates the image's pixel values and their spatial relationships. Ultimately, this complex serves as a mathematical foundation for analyzing and processing grayscale images.

We establish an algebraic structure within the context of the pixels' complex by defining k-chains, denoted as $C_k$, as formal sums of $k-$ cells. These $k$ -chains have coefficients in $\mathbb{Z}_2$, as orientation is not required for our purposes. The entire set of $k$ -chains, along with the addition operation, constitutes a group referred to as $C_k\,(X)$, and we call them the chain groups of dimension $k$. To maintain relationships between chains of varying dimensions, we employ boundary operators. The boundary operator $\delta_k$ is a function between chain groups and is denoted as $\delta_k$: $C_k\,(X) \rightarrow C_{k-1}\,(X)$. A chain complex is a sequence of chain groups connected by boundary functions, ensuring that $\delta_k \delta_{k+1} = 0$ holds for all dimensions k. In the case of 2D, the chain complex is represented as follows:

$$\emptyset \rightarrow \mathrm{C}_2 \rightarrow \mathrm{C}_1 \rightarrow \mathrm{C}_0 \rightarrow \emptyset \quad (1)$$

Where $\emptyset$ is the empty set. It's worth noting that this chain complex's extension to higher dimensions is straightforward. Furthermore, the crucial property that the boundary of a boundary equals emptiness plays a significant role, as it is essential for constructing homology groups, as we will explore further in the subsequent sections.

#### *Filtration of the pixels' complex*

Homology groups, as a concept, play a crucial role in our discussion before investigating the filtration scheme. These groups are computed based on the topological complex. The primary aim of homology groups is to identify and retain chains that encompass holes within the complex while disregarding those that do not. To achieve this goal, we categorize chains into two distinct groups. The first group comprises chains whose boundaries are null, it means those chains that give a null result when we apply the boundary function to them, and these are of particular significance in our construction. We refer to this group as the $k$-th cycle subgroup within $C_k$. Its importance lies in its ability to represent chains that form closed loops or cycles within the complex, which directly relates to the topological features of interest. We will note this subgroup $Z_k$.

$$Z_k = \{x \in C_K | \delta_k(x) = 0\} = ker\delta_k \quad (2)$$

Where $ker\delta_k$ is the kernel of the boundary function. Within these groups, our focus shifts to those elements that represent boundaries of higher-dimensional chains. This subset forms another group known as the $k$ -boundary group:

$$B_k = \{x \in C_K | \exists y \in \backslash C_{k+1}, x = \delta_{k+1} y\} = im\,\delta_{k+1} \quad (3)$$

Where $\mathrm{im}\delta_{\mathrm{k+1}}$ represents the image of the boundary function. As an example, the 1-chains B and G that are composed from blue and green edges in Fig. 2 are cycles because their boundaries are null. Indeed

$$\begin{aligned}\delta_1(\text{“Blue”}) = \delta_1((\mathrm{bc}) + (\mathrm{cd}) + (\mathrm{dh}) + (\mathrm{hl}) \\ + (\mathrm{ki}) + (\mathrm{if}) + (\mathrm{fb})) = \mathrm{b} \\ + \mathrm{c} + \mathrm{c} + \mathrm{c} + \mathrm{d} + \mathrm{d} + \mathrm{h} + \mathrm{h} \\ + \mathrm{l} + \mathrm{k} + \mathrm{i} + \mathrm{i} + \mathrm{f} + \mathrm{f} + \mathrm{b} = 0\end{aligned}$$

since every point appears twice and we work on $\mathbb{Z}_2$ field. Thus "$Blue$" belongs to $Z_1$. The 1-chain $B$ is also a 1-cycle and at the same time it is the boundary of a 2-chain, then "$Blue$" $\in B_1$. Due to the property $\delta_k \delta_{k=1} = 0$, it follows that the $k$ -boundary group $B_k$ is a subset of the $k$ -cycle group $Z_k$. As a result, we can define the $k$ -th homology group $H_k$ as the quotient group obtained by dividing $Z_k$ by $B_k$. The fundamental purpose of the homology group is to filter out cycles that are themselves boundaries since such cycles cannot enclose voids or holes. Consequently, our focus primarily lies on the $k$ -th homology group $H_k$, while the $k$ -boundary group $B_k$, which represents cycles that are boundaries, is reduced to a mathematical zero. Furthermore, since $H_k$ is a quotient group, it can be represented using equivalence classes. Each element of $H_k$ belongs to an equivalence class denoted as $[z]$, where $z$ is an element of $Z_k$. This class serves as a container for equivalent cycles, and every chain that is part of $Z_k$ but not a part of $B_k$ can be effectively represented within this equivalence class. In this way, $H_k$ captures the essential topological features of interest within the complex. The concept of a filtration scheme, coupled with the functionality of homology, leads to the emergence of the concept known as "homological persistence". This filtration scheme is established based on an intensity function, which serves as a criterion for ordering the cells within the complete complex X, encompassing all cells, based on their respective intensity values.

In our filtration structure, we assign the 0-cells to store pixel intensities as values, while each $k$ -cell stores the maximum value among its $k-1$ -cells, which represent its boundary. For example, a square cell will contain the maximum intensity value among its four edges, and so on. Following this approach, we denote $X_i$ as the pixels' complex, which contains cells with values not exceeding the integer $i$. Consequently, we create a nested sequence of subcomplexes within $X$, as follows:

$$\emptyset \subset X_0 \subset X_1 \subset \ldots \subset X_i \subset \ldots \subset X \quad (4)$$

Figure 3 provides an illustrative example of this filtration process applied to the pixels' complex from Fig. 2. As we progress from $X_0$ to $X_{27}$, new cells are gradually introduced, and cycles and boundary groups are formed during the filtration process. By tracking the topological changes in this filtration using homology, we obtain a sequence of homology groups for each dimension $k$, connected by linear maps induced by inclusions:

$$H_k(X_0) \rightarrow H_k(X_1) \rightarrow \ldots H_k(X_i) \rightarrow \ldots \rightarrow H_k(X) \quad (5)$$

This sequence of homology groups enables us to analyze the evolving topological characteristics of the complex as we transition from one filtration step to the next.

#### *Persistent homology computation*

We can identify homology classes at each level $X_i$ during the filtration process. However, this approach results in a loss of specific information regarding individual cycles. A more informative approach involves detecting the lifespans of homology classes and tracking their evolution throughout the filtration, rather than merely

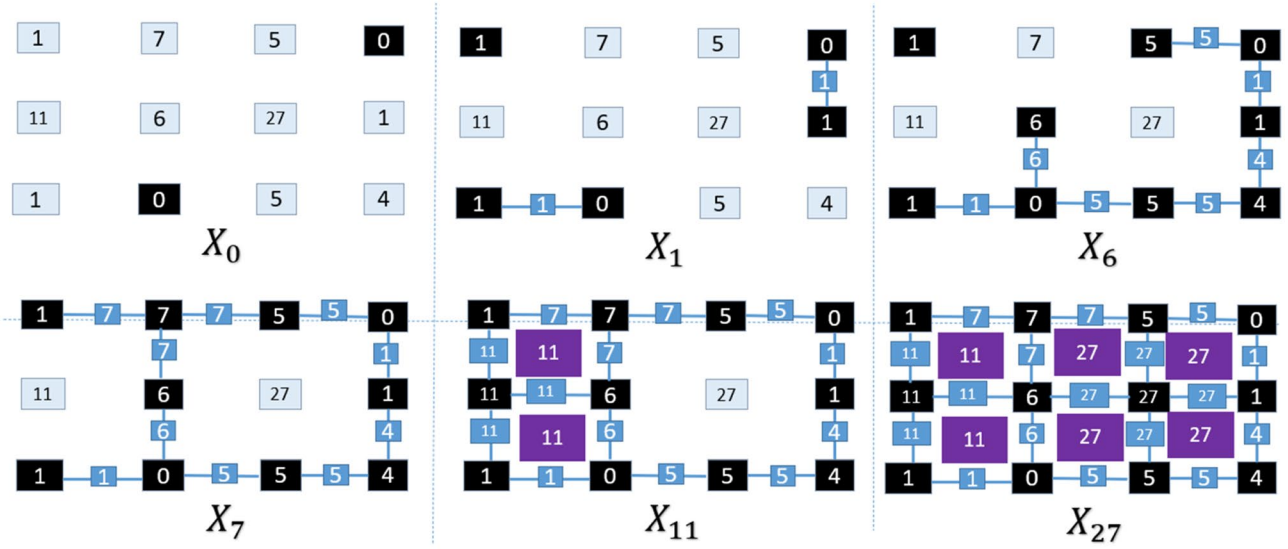


**Fig. 3** Example of filtration scheme of the pixels complex in Fig. 2

recording their existence at a single level $i$.During the intensity-based filtration, new homology classes may emerge, or existing ones may cease to exist at each level $i$. The lifespan of a homology class is determined by the difference between the level at which it is created and the level at which it is destroyed. For example, in Fig. 3, a 1-cycle representing a homology class is born at $X_7$ but later ceases to exist at $X_{27}$ because it becomes part of a higher-dimensional chain. Consequently, the lifespan of this cycle is 20. Homology classes with longer lifespans signify the presence of significant topological features in the image, while those with shorter lifespans are typically considered as noise. This observation can be visually represented using a persistence diagram, which records the birth and death times of homology classes. This perspective is further reinforced by the stability of the persistence diagram under continuous changes [35], making it particularly valuable in biomedical image processing, as it allows for the extraction of features that remain invariant to geometric transformations.

#### Persistent homology and machine learning

Persistent homology-based machine learning (PHML) models have found applications across various domains, including image analysis. The core concept behind PHML is to derive topological insights from data through persistent homology (PH) and subsequently integrate these features with machine learning techniques, encompassing both supervised and unsupervised learning methods [38]. Persistent homology can ensure several theoretical benefits when integrated with other methods. It focuses on the persistent topological features of the data, which are less sensitive to noise compared to geometric or algebraic methods. This robustness can be particularly useful when dealing with noisy or complex datasets [35, 39]. Persistent homology captures global structural information about the dataset, such as loops, voids, and tunnels, which may not be easily discernible using other methods. This ability to capture global structure can lead to more holistic representations of the data [33, 40]. One of the key advantages of persistent homology is its ability to analyze data across multiple spatial scales simultaneously. This multiscale analysis can uncover hierarchical structures and relationships in the data, providing insights that may be missed by methods that focus on a single scale [41, 42]. In the context of computational implementation (as depicted in Fig. 4), PHML can be divided into four key steps: cubical complex construction, PH analysis, extraction of topological features, and the utilization of topology-based machine learning. In one instance, as described in [43], the authors introduce the WSI-GTFE method, which combines Topological Data Analysis (TDA) and Graph Neural Networks (GNNs) to effectively identify and quantify key pathogenic information pathways, capturing both macro and micro architectural details within histology images. In another study [44] the authors explore an approach that enhances modern image classification techniques by incorporating topological features. This method has demonstrated impressive accuracy in image classification using deep learning algorithms. Furthermore, a recent development [45] introduces a novel approach that merges TDA and deep learning features for the purpose of COVID-19 detection from chest X-ray images.

The initial step involves the construction of the pixels' complex based on the data under investigation. Following this, in the second step, we employ filtration as part of persistent homology (PH) analysis. Through a carefully

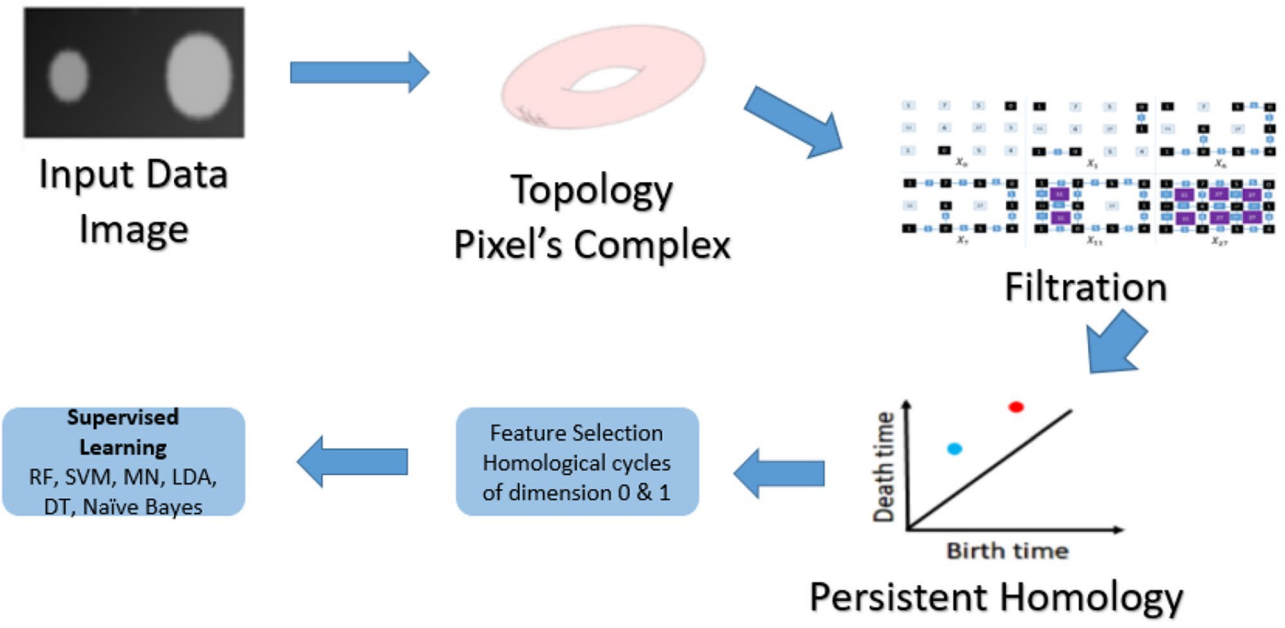


**Fig. 4** The flowchart for PHML modeling. It entails the selection of specific cubical complexes based on the data types in use. By tuning an appropriate filtration parameter, one can apply persistent homology (PH) analysis through custom software such as PHAT. The output generated by PH is then converted into feature vectors, which are subsequently integrated with supervised machine learning methods

designed filtration process, we can compute the persistence of topological properties. In simple terms, persistence provides information about the geometric extent of these properties.

Various software tools are available for conducting persistent homology analysis on diverse data structures. In our case, we utilize PHAT, which stands for the Persistent Homology Algorithm Toolbox, as introduced, and detailed in reference [46]. PHAT offers an efficient means of calculating persistent homology classes specifically tailored for images. Moving on to the third step, we focus on the extraction of meaningful topological features from the results obtained through persistent homology analysis. These results are typically represented in the form of either persistent barcodes (PBs) or persistence diagrams (PDs). Consequently, we convert the results into topological feature vectors from the PH results.

The final step entails combining these topological features with machine learning algorithms. In essence, these features can be directly employed in supervised learning models. Depending on the specific learning models chosen, it is essential to consider various feature selection and representation techniques to maximize the model's performance. For more detailed information, please refer to article [47].

#### Proposed methodology

In image segmentation, the topological characteristics derived from persistent homology play a significant role [48–50] Concerning our work, upon computing persistent homology within images, we can extract essential information such as the lifespan of 0-cycles, 1-cycles, and the measure of persistent entropy. In addition to these topological attributes, we also compute the average and standard deviation of the lifespan for 0-cycles and 1-cycles within each window, along with their respective persistent entropies for dimensions 0 and 1. Furthermore, we calculate the mean and standard deviation of pixel values, resulting in a comprehensive set of eight features computed within each image. The persistent entropy H of the topological space is calculated as follows:

$$H = -\sum_{i \in I} p_{\mathrm{i}} \log\left(p_{\mathrm{i}}\right) \tag{6}$$

Where $l_i$ represents the lifespans of the homology cycles, $L = \sum_{i \in I} l_{\mathrm{i}}, p_{\mathrm{i}} = \frac{l_{\mathrm{i}}}{L}$.

To summarize, from each CT scan, we extract a collection of eight features, encompassing the mean and standard deviation of 0-cycle lifespans, the mean and standard deviation of 1-cycle lifespans, 0-persistent entropy, 1-persistent entropy, as well as the mean and standard deviation of pixel values. The subsequent phase involves integrating this dataset into supervised machine learning algorithms, specifically designed for classification tasks. Supervised learning aims to construct a compact model representing the relationship between predictor features and class labels' distribution. Subsequently, this model is employed to assign class labels to test samples, where predictor feature values

are available, but class label values are unobserved. The evaluation of the classifier typically relies on prediction accuracy, calculated as the proportion of correct predictions divided by the total number of predictions made [51]. In this application, we've opted for Decision Trees (DT) as our logic-based learning approach, this method doesn't necessitate making any initial statistical assumptions about the data's distribution. For statistical learning algorithms, we've employed the Random Forest Classifier (RF), Support Vector Machines (SVM), Naïve Bayes Classification (NBC) and Logistic regression (LG) [38, 47, 52]. For SVM we used the RBF kernel as kernel function. The likelihood of Naive Bayes classifier used is the probability of the predictor given class.

Note that the machine learning techniques applied on homological persistence features are very effective in the detection of the right class in these kinds of images, a previous article showed the effectiveness of the choice of these 8 characteristics [53]. Finally, we compute the confusion matrix to measure the accuracy of the five supervised machine learning methods used.

## Data collection and image preprocessing

Our primary objective was to assemble a dataset comprising chest CT scan images from both normal and COVID-19 patients and prepare this dataset for seamless use in subsequent phases. We collected a total of 800 axial chest CT images, evenly distributed between patients with and without COVID-19. These images underwent meticulous manipulation and processing to eliminate unwanted pixels, such as those originating from ribs, the heart, and blood vessels, while retaining the crucial lung pixels that hold the pertinent information in our dataset. To enhance the overall accuracy, we performed a series of image preprocessing steps.

- Image Resize and Normalization: Since the images in the dataset had varying sizes, the initial preprocessing step involved resizing all images to a standardized dimension of 128 × 128 pixels. Subsequently, all images were converted to a uniform grayscale color map [0 256] and then normalized to a [0 1] scale.Image Enhancement: Image enhancement plays a pivotal role in digital image processing, facilitating a more effective visual inspection and further analysis of image groups. One of the techniques employed for this purpose is contrast enhancement, which seeks to amplify image contrast while preserving the original image brightness. To accentuate the contrast between the entire lung region of interest in chest CT images and surrounding structures (such as bones, soft tissues, blood vessels, and the heart), we applied a contrast enhancement technique known as Gamma Correction (GC) [54]. GC is a histogram modification technique that utilizes a variable parameter γ. In our case, we utilized the GC method with $\gamma = 3$ for its exceptional ability to enhance the distinction between the lungs and other components in all dataset images. Noting that no additional noise reduction techniques were applied beyond the described preprocessing steps.Lung Segmentation: The primary goal of lung segmentation in our study was to retain only the pixels relevant to lung parenchyma, whether with or without COVID-19 findings, and to set all other pixels from surrounding structures to zero. This process was carried out to improve the accuracy of predictive models to be employed in subsequent sections. Lung segmentation involved several key steps, as outlined below:

  1. Image binarization was performed using binary and triangle thresholding methods for normal images [55] and the OTSU Thresholding method for COVID images [56].
  2. Image Mask Generation involved detecting all in-lung nodules and out-lung objects and subsequently removing them from the image [57].
  3. Applying the inversion of the generated mask onto the original image yielded the segmented lung image, which exclusively retained lung pixels while eliminating all other extraneous pixels.

The Fig. 5 illustrates examples of chest CT scan images of normal and COVID-19 patients after performing the mentioned preprocessing steps.

## Evaluation metrics

We divided our dataset into training and testing subsets, varying the testing sizes. The subsequent results illustrate the performance of our supervised algorithms. For each approach, we computed the confusion matrix, accuracy, F1 score, and AUC score. The diagonal elements in a confusion matrix represent correct classifications, while off-diagonal elements represent misclassifications. It consists of four key elements beginning from first row, first column, True Positives (TP), False Negatives (FN), False Positives (FP), and True Negatives (TN). In addition to this, cross-validation is widely employed in applied machine learning to test the efficacy of a model on unseen data. This technique involves utilizing a portion of the dataset for training and the rest for validation, ensuring a more realistic assessment of model performance. Cross-validation, specifically the k-fold method, with $k = 5$ is used in our method [58]. This approach partitions the dataset into k subsets, allowing for robust validation and testing. Each subset is used iteratively for

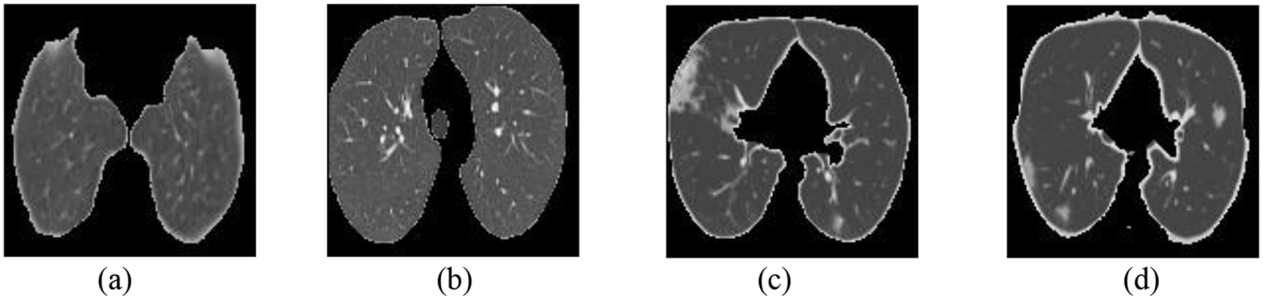


**Fig. 5** Images (**a**) and (**b**) illustrate chest CT scans of individuals without any abnormalities, whereas images (**c**) and (**d**) represent scans of patients diagnosed with COVID-19

**Table 2** Decision Tree Classifier results

| Test size | Accuracy (%) | F1 score | AUC score | Precision % | Recall% | Specificity % | Confusion Matrix |
|---|---|---|---|---|---|---|---|
| 0.1 | 91.25% | 0.8889 | 0.90090 | 96.55% | 82.35% | 97.83% | 28 6<br>1 45 |
| 0.2 | 93.75% | 0.9419 | 0.93812 | 93.10% | 95.29% | 92.00% | 81 6<br>4 69 |
| 0.3 | 90.42% | 0.9084 | 0.90382 | 91.20% | 90.48% | 90.35% | 114 11<br>12 103 |
| 0.4 | 90.63% | 0.9112 | 0.90551 | 94.48% | 88.00% | 93.79% | 154 9<br>21 136 |

validation while the remaining data is used for training, with the process repeated multiple times for thorough evaluation.

In Table 2, we present the outcomes obtained using the Decision Tree Classifier. We calculated the confusion matrix, which yielded an accuracy of 93.75% for a 20% testing size, along with an F1 score of 0.94. The AUC score achieved for this classifier was 0.93812. The values of precision, recall and specificity are mentioned as well in the table. In addition, we used the same k-fold cross validation with k = 3,4,5 and 10. The mean and standard deviation of accuracy results across splits were 91.72% and 1.31% respectively which shows the consistency of the results.

Moving on to Table 3, we showcase the results obtained from employing the Random Forest Classifier. The confusion matrix was displayed, showing an accuracy of 97.51% for a 30% testing size, and an AUC score above 0.99 based on the ROC curve for all testing sizes. The mean and standard deviation of accuracy results across splits, for k = 3,4,5 and 10, were 96.34% and 0.72% respectively which shows a high consistency.

In Table 4, we present the findings from applying the Naïve Bayes Classifier. We implemented the confusion matrix and computed the average accuracy, F1 score, and AUC score, which were 84.09%, 0.8559, and 0.948845, respectively It's worth noting that this method exhibited relatively lower performance compared to the other classification models. The mean and standard deviation of accuracy results across splits, for k = 3,4,5 and 10, were 84.11% and 1.64% respectively.

In Table 5, we exhibit the results derived from the Support Vector Machine Classifier. We displayed the confusion matrix, revealing an accuracy of 96.25% for a 30% testing size, and an AUC score of 0.992495 based on the ROC curve. Notably, the SVM classifier achieved the highest AUC value (0.992495) among all the methods, surpassing the others (DT: 0.93812, RF: 0.990278, NB: 0.95386 LG: 0.945505). Consequently, it stands out as the most efficient method for detecting COVID-19 in CT-scan images. The mean and standard deviation of accuracy results across splits, for k = 3,4,5 and 10, were 94.51% and 0.95% respectively.

Lastly in Table 6, we showcase the results obtained from employing the Logistic regression Classifier. The confusion matrix was displayed, showing a highest accuracy of 85.42% for a 30% testing size, and an average AUC score of 0.945505 based on the ROC curve. The mean

**Table 3** Random Forest Classifier results

| Test size | Accuracy (%) | F1 score | AUC score | Precision % | Recall% | Specificity % | Confusion Matrix |
|---|---|---|---|---|---|---|---|
| 0.1 | 96.25% | 0.9552 | 0.99213 | 96.97% | 94.12% | 97.83% | 32 1<br>2 45 |
| 0.2 | 95.63% | 0.9600 | 0.99212 | 94.38% | 97.67 | 93.24% | 84 5<br>2 69 |
| 0.3 | 97.51% | 0.9750 | 0.99128 | 96.67% | 0.9831 | 96.72% | 116 4<br>2 118 |
| 0.4 | 95.94% | 0.9590 | 0.99236 | 97.44% | 94.41% | 97.48% | 152 4<br>9 155 |

**Table 4** Naïve Bayes Classifier results

| Test size | Accuracy (%) | F1 score | AUC score | Precision % | Recall% | Specificity % | Confusion Matrix |
|---|---|---|---|---|---|---|---|
| 0.1 | 83.75% | 0.8571 | 0.95301 | 92.86% | 79.59% | 90.32% | 39 3<br>10 28 |
| 0.2 | 81.88% | 0.8304 | 0.95019 | 95.95% | 73.19% | 95.24% | 71 3<br>26 60 |
| 0.3 | 86.67% | 0.8750 | 0.95386 | 95.73% | 80.58% | 95.05% | 112 5<br>27 96 |
| 0.4 | 84.06% | 0.8610 | 0.93832 | 94.43% | 79.70% | 90.98% | 158 11<br>40 111 |

**Table 5** Support Vector Machine Classifier results

| Test size | Accuracy (%) | F1 score | AUC score | Precision % | Recall% | Specificity % | Confusion Matrix |
|---|---|---|---|---|---|---|---|
| 0.1 | 93.75% | 0.9315 | 0.98921 | 97.14% | 89.47% | 97.62% | 34 1<br>4 41 |
| 0.2 | 94.38% | 0.9508 | 0.99022 | 98.86% | 91.58% | 98.46% | 87 1<br>8 64 |
| 0.3 | 96.25% | 0.9639 | 0.99350 | 97.56% | 95.24% | 97.37% | 120 3<br>6 111 |
| 0.4 | 94.69% | 0.9489 | 0.98586 | 97.53% | 92.40% | 97.48% | 158 4<br>13 145 |

**Table 6** Logistic Regression Classifier results

| Test size | Accuracy (%) | F1 score | AUC score | Precision % | Recall% | Specificity % | Confusion Matrix |
|---|---|---|---|---|---|---|---|
| 0.1 | 83.75% | 0.8222 | 0.94561 | 90.24% | 80.43% | 88.24% | 37 4<br>9 30 |
| 0.2 | 82.50% | 0.8313 | 0.94352 | 95.83% | 73.40% | 95.45% | 69 3<br>25 63 |
| 0.3 | 85.42% | 0.8606 | 0.95105 | 93.10% | 80.00% | 92.38% | 108 8<br>27 97 |
| 0.4 | 83.75% | 0.8556 | 0.94184 | 90.06% | 81.48% | 86.95% | 154 17<br>35 114 |

and standard deviation of accuracy results across splits, for k = 3,4,5 and 10, were 83.86% and 1.03% respectively.

## Discussion

Based on the obtained results, the Random Forest Classifier demonstrated the highest level of accuracy compared to the other classifiers. Specifically, it achieved an accuracy rate of 97.5%, surpassing the Decision Tree Classifier at 93.75%, the Naïve Bayes Classifier at 86.7%, and the Support Vector Machine at 96.25% and the Logistic Regression at 85.42%. When it comes to the Area Under the Curve (AUC) metric, the Support Vector Machine (SVM) classifier outperformed all other methods, with an AUC value of 0.992495 which exceeded those of the Decision Tree (0.93812), Random Forest (0.990278),Naïve Bayes (0.95386) and Logistic Regression (0.945505).

When choosing classifiers, several factors come into play to identify the most suitable models. Decision Trees are straightforward to interpret and understand, offering a clear depiction of decision rules. This clarity makes them particularly valuable in situations where model explainability is crucial. Random Forest, on the other hand, is an ensemble method that combines multiple Decision Trees to enhance generalization and robustness. By averaging the predictions from a collection of trees, it reduces variance and minimizes the risk of overfitting that can occur with a single tree. This method can also assist in identifying important features, which is beneficial for feature selection or for gaining insights into which features hold more significance in the data. SVMs are adept at managing high-dimensional spaces and perform well with high-dimensional datasets. However, they can be computationally demanding, particularly with large datasets, and may necessitate careful tuning of the kernel, regularization, and margin parameters. It's important to note that accuracy can be sensitive to the threshold used to convert predicted probabilities into class labels. Altering this threshold may result in different accuracy values. In contrast, the AUC-ROC provides a more comprehensive performance measure that is not dependent on the threshold. The remarkable performance of SVM can likely be attributed to the incorporation of topological features obtained through operations that engage linear equations [38, 59]. This choice aligns well with the superior performance of the Support Vector Machine, which excels when dealing with linearly separable data [52]. The Naïve Bayes classifier leverages probabilistic principles, while the Decision Tree and Random Forest Classifier rely on tree-like structures for decision-making. In addition, we remark that the Random Forest Classifier accuracy is the highest in all testing sizes between all the classification methods and the Support Vector Machine method gives stable AUC results that are the highest in a testing size of 30% which represents a balance between precision and generalization. In addition, a kernel-based method on the raw distances, like SVM is applied directly on the images since it's the most popular type of kernel approach. A binary classifier determines the best hyperplane that most effectively divides the two groups. To efficiently locate the ideal hyperplane, SVMs map the input into a higher-dimensional space using a kernel function. This method showed accuracy with less than 90% for all testing sizes.

In order to assess how much better of a classifier we can build if we incorporate topological features to the analysis, we applied the exact same models that we fitted, but with just the pixel average and standard deviation as inputs. The classification results showed a maximum accuracy of about 94.7% and a maximum AUC value of 0.97328. In addition, these same models were applied on the 6 topological features alone and showed a maximum accuracy of 85.4% and a maximum AUC value of 0.86519. These results show clearly that the topological features combined with the statistical ones that form the 8 features improve the accuracy percentage of the classification methods. The use of topological data analysis

provides a new perspective by revealing significant structural patterns in image data that existing methods may overlook. Topological features allow the model to identify and assess the connectivity and gaps within CT images. These features reflect the essential geometry and topology of lung tissues, which can change due to infection. In cases of COVID-19, affected lungs often exhibit unique characteristics such as ground-glass opacities, consolidations, and fine reticular patterns that disrupt the natural topology of lung structures. By integrating these topological features, it becomes easier to distinguish between different patterns in CT images that correspond to various stages and severity of COVID-19. These topological changes are captured by the persistence of certain features across multiple scales, enhancing the model's sensitivity and specificity.

In addition, the same models were applied on CT scan images from the dataset used in [30] to validate our method on a different dataset. The results indicate that both the SVM and Random Forest classifiers outperformed the others, displaying notably high accuracy levels. Specifically, the SVM achieved an accuracy rate of 94.5% with an AUC score of 0.93, while the Random Forest classifier achieved an accuracy rate of 93.8% and an AUC score of 0.94. These findings underscore the effectiveness of the proposed method when applied to additional datasets of CT scan images of lung images.

To contextualize the novelty and effectiveness of the proposed approach, we applied deep learning and conventional image processing techniques on the data set. The results of the three and four convolutional layers CNN show a best accuracy of 94.2 % and 94.8% respectively with Decision Tree classifier, 92.3% and 93.4% with Random Forest classifier, 93.4% and 94.1% with Naïve Bayes classifier, 95.1% and 95.7% with SVM classifier.

The study may encounter limitations due to the variability in CT image quality, which is affected by factors such as resolution, noise, and contrast that vary between different equipment. This inconsistency could reduce the model's accuracy when it is applied to real-world clinical data, particularly with low-quality images that may have motion artifacts or poor resolution, leading to potential false positives or negatives. Furthermore, while the use of topological features is innovative, it may present challenges for clinical interpretability, as these features do not provide the intuitive insights that traditional imaging biomarkers offer.

## Conclusion

Numerous studies in the field of image analysis have harnessed the power of persistent homology and the associated statistical insights to categorize image data. In our research, we leveraged the results of persistent homology to distinguish between COVID and non-COVID images. This endeavor unveils a groundbreaking detection method characterized by its parameter-free nature, thanks to the innovative concept of persistent homology profiles. When it comes to the classification of biomedical images, the fusion of machine learning techniques with persistent homology is highly recommended due to its effectiveness [53, 60, 61]. In particular, topological features derived from biomedical images, notably from CT scans, are portrayed in terms of the scale and distribution of zero and one-dimensional persistence features through the lens of persistent homology. We introduce a unified platform designed for the automatic detection of COVID-19 by applying machine learning techniques to these features. The Random Forest classifier demonstrated an impressive 97.5% accuracy in COVID-19 detection, while the SVM classifier exhibited an exceptional AUC of 0.992495, surpassing the performance of existing state-of-the-art methods applied to the task of COVID-19 detection in CT scan images. A comparison made on the same data by using only statistical features showed the efficiency of adding topological features for the classification. To validate the method, the same model was applied on other chest CT scan images showing high accuracy. These findings highlight the significant potential of our approach and underscore the benefits of integrating topological image characteristics into classification tasks.

## Future work

While this is a preliminary work and yields no definitive conclusions regarding clinical utility, it does point to the power of persistent homology based on machine learning, to interrogate the architectures present in COVID-19. Using this tool, it may be possible to develop a better understanding of how it correlates with patient prognostic outcomes. There are many directions for further research. On the processing front, it is crucial to thoroughly assess the effectiveness of our findings in characterizing the severity and progression of COVID-19, enabling physicians to make better-informed decisions. For this aim, the dataset diversion needs to be increased. The machine learning methods should be implemented more on other CT images, X-ray chest images. A model based on coarse summaries of the data that requires statistical tools like quartiles can be applied to investigate the results. Other aspects may cover the investigation of the order of importance of the 8 summaries found before classification. Principal component analysis (PCA) or Variational Autoencoder methods can be applied for this purpose.

#### Acknowledgements

This study was carried out by Holy Spirit University of Kaslik, Faculty of Arts and Sciences in partnership with University of Reims Champagne Ardenne in France

**Author contributions**
All named authors contributed equally to the construction of the paper. RA designed the structure of this article and managed to apply topological concepts on the images. AR run the classification algorithms and interpreted the results. MK contributed in the explanation of mathematical methods and discussion of the results. He also reviewed the article for faults and added some other explanations and also revised the manuscript for linguistic check and some other explanations. AG and VV were responsible of illustration part and preprocessing of images. Authors have read and agreed to the published version of the manuscript.

**Funding**
This project has been funded with the support of the Holy Spirit University of Kaslik.

**Data availability**
All data supporting the findings of this study are available under request. Examples of the data are showed in the manuscript. We ensure that the Availability of Data and Materials statement exactly matches the main manuscript and submission system.

## Declarations

**Ethics approval and consent to participate**
The need for ethics approval is deemed unnecessary according to national regulations, no relevant legislation exists in the country of the origin of the data. The data was anonymized prior to use, and all reasonable steps were taken to ensure confidentiality and adherence to ethical standards for research involving medical data.

**Accordance**
As clinical/pathological data is analyzed in this study, we confirm that all methods were performed in accordance with the Declaration of Helsinki.

**Consent for publication**
Not applicable

**Competing interests**
The authors declare no competing interests.

Received: 9 April 2024 / Accepted: 18 February 2025
Published online: 02 April 2025

**References**

1. Ciotti M, et al. The COVID-19 pandemic. Crit Rev Clin Lab Sci. 2020;57:365–88.
2. Shi Y, et al. An overview of COVID-19. J Zhejiang Univ Sci B. 2020;21:343–60.
3. Peaper DR, Kerantzas CA, Durant TJS. Advances in molecular infectious diseases testing in the time of COVID-19. Clin Biochem. 2023;117:94–101.
4. Stolberg-Stolberg J, et al. COVID-19 rapid molecular point-of-care testing is effective and cost-beneficial for the acute care of trauma patients. Eur J Trauma Emerg Surg. 2023;49:487–93.
5. Hayden MK, et al. The infectious diseases Society of America Guidelines on the Diagnosis of COVID-19: antigen Testing. Clinl Infect Dis. 2023. https://doi.org/10.1093/cid/ciad032.
6. Widyasari K, Kim S. Rapid Antigen Tests during the COVID-19 Era in Korea and their implementation as a detection tool for other infectious diseases. Bioengineering (Basel). 2023;10:322.
7. Sullivan D, Casadevall A. COVID-19 Serology data provide guidance for future deployments of Convalescent Plasma. mBio. 2023;14:e0042823–e0042823.
8. Sunnetci KM, Alkan A. Biphasic majority voting-based comparative COVID-19 diagnosis using chest X-ray images. Expert Syst Appl. 2023;216:119430.
9. Das D, Biswas SK, Bandyopadhyay S. Perspective of AI system for COVID-19 detection using chest images: a review. Multimed Tools Appl. 2022;81:21471–501.
10. Yildirim M, Eroğlu O, Eroğlu Y, Çinar A, Cengil E. COVID-19 detection on Chest X-ray images with the proposed model using artificial intelligence and classifiers. New Gener Comput. 2022;40:1077–91.
11. Aslani S, Jacob J. Utilisation of deep learning for COVID-19 diagnosis. Clin Radiol. 2023;78:150–57.
12. Ullah F, Moon J, Naeem H, Jabbar S. Explainable artificial intelligence approach in combating real-time surveillance of COVID19 pandemic from CT scan and X-ray images using ensemble model. J Supercomput. 2022;78:19246–71.
13. Dairi A, Harrou F, Zeroual A, Hittawe MM, Sun Y. Comparative study of machine learning methods for COVID-19 transmission forecasting. J Biomed Inform. 2021;118:103791.
14. Pinter G, Felde I, Mosavi A, Ghamisi P, Gloaguen R. COVID-19 Pandemic Prediction for Hungary; A Hybrid Machine Learning Approach. Mathematics. 2020;8:890.
15. Aboughazala L. Automated detection of Covid-19 coronavirus cases using deep neural networks with X-ray images. J Med Virus Res Stud. 2020;2:1–12.
16. Wang L, Lin ZQ, Wong A. COVID-Net: a tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images. Sci Rep. 2020;10:19549.
17. Karar ME, Hemdan EE-D, Shouman MA. Cascaded deep learning classifiers for computer-aided diagnosis of COVID-19 and pneumonia diseases in X-ray scans. Complex Intelligent Syst. 2021;7:235–47.
18. Narin A, Kaya C, Pamuk Z. Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks. Pattern Anal Appl. 2021;24:1207–20.
19. Khanday AMUD, Rabani ST, Khan QR, Rouf N, Mohi Ud Din M. Machine learning based approaches for detecting COVID-19 using clinical text data. Int J Inf Technol. 2020;12:731–39.
20. Dutta P, Paul S, Kumar A. Comparative analysis of various supervised machine learning techniques for diagnosis of COVID-19. In: electronic devices, circuits, and systems for biomedical applications: challenges and intelligent approach. Elsevier; 2021. pp. 521–40. https://doi.org/10.1016/B978-0-323-85172-5.00020-4
21. Yao H, et al. Severity detection for the coronavirus disease 2019 (COVID-19) patients using a machine learning model based on the blood and urine tests. Front Cell Dev Biol. 2020;8:683.
22. Tamal M, et al. An integrated framework with machine learning and radiomics for accurate and rapid early diagnosis of COVID-19 from Chest X-ray. Expert Syst Appl. 2021;180:115152.
23. Brinati D, et al. Detection of COVID-19 infection from routine blood exams with machine learning: a feasibility study. J Med Syst. 2020;44:135.
24. Amézquita EJ, Quigley MY, Ophelders T, Munch E, Chitwood DH. The shape of things to come: topological data analysis and biology, from molecules to organisms. Dev Dyn. 2020;249:816–33.
25. Grommé F, Ruppert E. Population geometries of Europe: the topologies of data cubes and grids. Sci Technol Hum Values. 2019;45:235–61.
26. Pham TD. A comprehensive study on classification of COVID-19 on computed tomography with pretrained convolutional neural networks. Sci Rep. 2020;10:16942.
27. Shah V, et al. Diagnosis of COVID-19 using CT scan images and deep learning techniques. Emerg Radiol. 2021;28:497–505.
28. Pathak Y, et al. Deep transfer learning based classification model for COVID-19 disease. Ing Rech Biomed. 2022;43:87–92.
29. Tiwari RS, D L, Das TK, Srinivasan K, Chang C-Y. Conceptualising a channel-based overlapping CNN tower architecture for COVID-19 identification from CT-scan images. Sci Rep. 2022;12:18197.
30. Sinha A, Joshi SP, Das PS, Jana S, Sarkar R. An ML prediction model based on clinical parameters and automated CT scan features for COVID-19 patients. Sci Rep. 2022;12:11255.
31. Subramanian M, Sathishkumar VE, Cho J, Shanmugavadivel K. Learning without forgetting by leveraging transfer learning for detecting COVID-19 infection from CT images. Sci Rep. 2023;13:8516.
32. Lee EH, et al. Deep COVID DeteCT: an international experience on COVID-19 lung detection and prognosis using chest CT. NPJ Digit Med. 2021;4:11.
33. Edelsbrunner H, Harer J. Computational Topology. American Mathematical Society. 2009. https://doi.org/10.1090/mbk/069.
34. Zomorodian A, Carlsson G. Computing Persistent Homology. Discrete Comput Geometry. 2004;33:249–74.
35. Cohen-Steiner D, Edelsbrunner H, Harer J. Stability of persistence diagrams. Discrete Comput Geometry. 2006;37:103–20.
36. Rieck B, et al. Uncovering the topology of time-varying FMRI data using cubical persistence.
37. Kaczynski T, Mischaikow K, Mrozek M. Computational Homology. 2004;157.
38. Hensel F, Moor M, Rieck B. A survey of topological machine learning methods. Front Artif Intell. 2021. vol. 4 Preprint at https://doi.org/10.3389/frai.2021.681108.

39. Jesper A, Møller M, Collins BM. Hand-in Information Titel: persistent Homology and Noise Må Besvarelsen Gøres Til Genstand for Udlån: yes.
40. Edelsbrunner H, Morozov D Persistent homology: theory and practice.
41. Yoon HR, Ghrist RW Persistence by parts: multiscale feature detection via distributed persistent homology.
42. Choi H, et al. Abnormal metabolic connectivity in the pilocarpine-induced epilepsy rat model: a multiscale network analysis based on persistent homology. Neuroimage. 2014;99:226–36.
43. Levy J, Haudenschild C, Barwick C, Christensen B, Vaickus L. Topological feature extraction and visualization of whole slide images using graph neural networks. 2020;16. www.worldscientific.com
44. Dey T, Mandal S, Varcho W Improved image classification using topological persistence. in Proceedings of the conference on vision, modeling and visualization. Goslar, DEU, Eurographics Association; 2017. pp. 161–68. https://doi.org/10.2312/vmv.20171272.
45. Hajij M, Zamzmi G, Batayneh F. TDA-Net: fusion of Persistent Homology and Deep Learning Features for COVID-19 Detection From Chest X-Ray Images. Annu Int Conf IEEE Eng Med Biol Soc. 2021;2021:4115–19.
46. Bauer U, Kerber M, Reininghaus J, Wagner HP. Persistent Homology Algorithms Toolbox. J Symb Comput. 2017;78:76–90.
47. Pun CS, Xia K, Lee SX. Persistent-homology-based machine learning and its applications – a survey. SSRN Electron J. 2018. https://doi.org/10.2139/ssrn.3275996.
48. Qaiser T, et al. Persistent homology for fast tumor segmentation in whole slide histology images. Procedia Comput Sci. 2016;90:119–24.
49. Assaf R, Goupil A, Vrabie V, Boudier T, Kacim M. Persistent homology for object segmentation in multidimensional grayscale images. Pattern Recognit Lett. 2018;112:277–84.
50. Byrne N, Clough JR, Montana G, King AP. A persistent homology-based topological loss function for multi-class CNN segmentation of cardiac MRI. Stat Atlases Comput Models Heart. 2020;2020:3–13.
51. Jen L. A brief overview of the accuracy of classification algorithms for data prediction in machine learning applications. J Appl Data Sci. 2021;2:84–92.
52. Pattern Recognition and Machine Learning. Pattern Recognition and Machine Learning. 2006. https://doi.org/10.1007/978-0-387-45528-0.
53. Rammal A, Assaf R, Goupil A, Kacim M, Vrabie V. Machine learning techniques on homological persistence features for prostate cancer diagnosis. BMC Bioinf. 2022;23:476.
54. Kumbhar M. Image and Video Dehazing based on Gamma correction method contrast enhancement. SSRN Electron J. 2021. https://doi.org/10.2139/ssrn.3882572.
55. Chaki N, Shaikh SH, Saeed K. (Computer scientist). Exploring image binarization techniques. 82.
56. Xu X, Xu S, Jin L, Song E. Characteristic analysis of Otsu threshold and its applications. Pattern Recognit Lett. 2011;32:956–61.
57. Zhou Y, et al. Collaborative learning of semi-supervised segmentation and classification for medical images.
58. Xu Y, Kong X, Cai Z. Cross-validation strategy for performance evaluation of machine learning algorithms in underwater acoustic target recognition. Ocean Eng. 2024;299:117236.
59. Lin Y, Wang S, Zhao L, Wang DK. Topology potential-based parameter selecting for support vector machine. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2014;8933:513–22.
60. Krishnapriyan AS, Montoya J, Haranczyk M, Hummelshøj J, Morozov D. Machine learning with persistent homology and chemical word embeddings improves prediction accuracy and interpretability in metal-organic frameworks. Sci Rep. 2021;11:1–11.
61. Arfi B. The promises of persistent homology, machine learning, and deep neural networks in topological data analysis of democracy survival. Qual Quant. 2023. https://doi.org/10.1007/s11135-023-01708-6.


## Publisher's Note