Topo-CXR: Chest X-ray TB and Pneumonia Screening with Topological Machine Learning

Faisal Ahmed[1] Brighton Nuwagira[1] Furkan Torlak[2] Baris Coskunuzer[1]
[1] Dept. of Math. Sciences, UT Dallas, [2] Dept. of Radiology, UT Southwestern
{faisal.ahmed, brighton.nuwagira, coskunuz}@utdallas.edu, furkan.torlak@utsouthwestern.edu

Abstract

Examination of chest X-ray images is currently one of the most important methods for the screening and diagnosis of thoracic diseases and, in some cases, for assessing response to treatment. However, this task is time-consuming and expensive as it requires a detailed visual inspection and interpretation by a trained clinician. In the past decade, several machine learning (ML) methods have been developed to remedy this issue as clinical decision support methods. However, most of these algorithms face challenges like computational feasibility, reliability, and interoperability.

In this paper, we develop a unique feature extraction method for chest X-rays by applying the latest topological data analysis (TDA) methods. We observe that normal and abnormal images produce very distinct topological patterns for pneumonia and tuberculosis. By using cubical persistence, we capture these patterns and convert them into powerful feature vectors. By combining with standard ML methods, we obtain a computationally feasible and interpretable model. In our extensive experiments, our model Topo-CXR outperforms state-of-the-art deep learning (DL) models in several benchmark datasets. Unlike most DL models, our proposed Topo-CXR model does not need any data augmentation or pre-processing steps and works perfectly on small datasets. Furthermore, our topological feature vectors can be easily integrated with any future ML and DL models to boost their performance and improve robustness.

1. Introduction

With approximately 4 million deaths each year, Pneumonia is one of the leading causes of death in infants and the elderly. It can be caused by a virus, bacteria, or fungus that have a predilection for colonizing the host alveoli. Patients with underlying disorders like asthma and immune compromise, or fragile populations like hospitalized infants and geriatric patients on ventilators are at increased risk for poor outcomes.

Similarly, despite being a curable and preventable communicable disease, Tuberculosis (TB) was the leading cause of death from a single infectious agent ranking above HIV/AIDS before the emergency of COVID-19 hence making it one of the major causes of ill health according to a 2021 Global Tuberculosis report [43]. Apart from TB and Pneumonia, there are other pathologic thoracic (chest) diseases such as; Pleural Effusion, Mass, Pneumothorax, Pulmonary Edema, Emphysema, Pulmonary Fibrosis, Pleural Thickening, and Atelectasis.

Chest X-rays are one of the most common types of radiological examinations [11] for the initial assessment of these thoracic diseases. However, radiology involves decision making under conditions of uncertainty and therefore can not always produce reliable results [10]. Additionally, it is limited by a need for qualified staff to individually interpret every radiograph [40]. Therefore, there has always been a need for a comprehensive and automated method for detecting these diseases.

Over the past decades, scientific research has produced tremendous innovations in medical imaging; however, chest X-rays remain a staple in the diagnosis of thoracic diseases. The recent advances in machine learning (ML) and deep learning (DL) have a great potential to offer a solution for early diagnosis and treatment planning. However, because of various challenges like the restrictive size of the training dataset, long pre-processing times, the need for high performance computing architectures, and the lack of interpretability in decision making, most such ML and DL methods fail to reach clinical stage implementation to address the crucial need for an automated screening method in this domain.

In this paper, we present a novel approach for diagnosing thoracic diseases using chest X-ray images, leveraging the application of topological data analysis (TDA). Over the years, TDA tools have exhibited remarkable success in medical image analysis across various domains (refer to Section 2.2). The fundamental concept behind TDA involves capturing hidden shape patterns within images and generating reliable representations of these patterns. The

topological features extracted by TDA tools offer distinctive insights into the data, enabling the discovery of new information and facilitating the determination of the most relevant features for the desired outcome. When combined with suitable machine learning (ML) models, these potent topological feature vectors empower the development of ML models that are both interpretable and robust, while maintaining high performance.

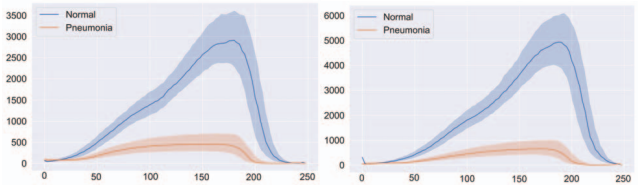
1.1. Our contributions

- We bring a novel perspective to chest X-ray screening with real world clinical utility by introducing the latest TDA methods to the field.
- By studying the evolution of topological patterns in chest X-rays, we observe that normal and abnormal images produce very distinct topological patterns for Pneumonia and TB. Employing cubical persistence, we obtain highly effective feature vectors capturing these distinct topological patterns (Figure 1 and 5).
- Our topological ML model gives outstanding results in detecting Pneumonia and TB outperforming the SOTA DL methods on benchmark datasets (Table 2, 3 and 4).
- Unlike various DL models, our model does not need any data augmentation and preprocessing and gives excellent results even in small datasets. Our model is computationally very fast - the end to end process for thousands of images takes only a few hours (Sec. 5.3).
- With our powerful topological descriptors, our proposed model is highly explainable and interpretable (Sec. 4.3).

2. Related Work

2.1. ML in Chest X-ray Analysis

The integration of machine learning techniques with medical image analysis has emerged as a groundbreaking paradigm, revolutionizing the field of healthcare diagnosis and treatment. With the exponential growth of available medical imaging data and the increasing complexity of diseases, traditional manual interpretation of images has become time-consuming and error-prone. In this context, machine learning algorithms offer immense potential by automating image analysis tasks, extracting relevant information, and aiding in accurate and timely decision-making. Since the development of Convolutional Neural Networks (CNNs), and its success in image classification in general [66], deep learning has become a desirable technique for most medical image analysis tasks due to its superior performance [15].



(a) Pneumonia vs. Normal (Betti-0) (b) Pneumonia vs. Normal (Betti-1)

Figure 1: We give the median curves and 40% confidence bands of our topological feature vectors (Betti functions) for each class in Ped-Pneumonia dataset. x-axis represents grayscale values and y-axis represents count of components (Betti-0) or count of loops (Betti-1) (Sec. 4.3).

For chest X-ray image analysis, there are four mainstream applications of ML tools [11]. The first is Image-level prediction networks which are used in tasks involving making prediction of a category label (classification) or a continuous value (regression). They are implemented by analyzing entire chest X-ray images. Some of the successful deep convolutional architectures for this category includes AlexNet [38], the VGG family of models [56], the ResNet family of models [29], DenseNet models [31], Inception Resnet [61], and the Xception network architecture [16]. The second mainstream application is the Segmentation networks, e.g U-Net [54]. The third mainstream application is the Localization networks, which include RCNN (Region Convolutional Neural Network) [24], and its variations [2]. Finally, the forth is Image generation networks, which include Generative Adversarial Network (GAN) [26]. In this paper, we focus only on the Image-level prediction networks to make predictions of chest X-ray diseases by using TDA methods.

2.2. TDA in Image Processing

Persistent homology, the main tool in TDA, has been quite effective for pattern recognition in image and shape analysis in the past two decade. There have been several works applying TDA in various fields of image analysis, e.g. for analysis of images of hepatic lesions [1], human and monkey fibrin images [7], fingerprint classification [22], ophthalmology [21, 19], analysis of 3D shapes [58], neuronal morphology [35], cancer detection/grading [48, 45, 17]. angiography [6], fMRI data [53, 60], and genomic data [12]. See the excellent survey [57] for a thorough review of TDA methods in biomedicine. Note also that TDA Applications Library [25] presents hundreds of interesting applications of TDA in various fields.

3. Background on TDA

In this paper, we use persistent homology (PH) as a powerful feature extraction tool for chest X-ray images. PH is one of the key approaches in topological data analysis (TDA), allowing us to systematically assess the evolution

of various hidden patterns in the data as we vary a scale parameter [70, 14]. The extracted patterns, or homological features, along with information on how long such features persist throughout the considered filtration of a scale parameter, convey a critical insight into salient data characteristics and data organization. Here, we give a basic introduction to PH in image setting (*cubical persistence*) for non-experts. For a more thorough background and PH process for other data types (e.g., point clouds, networks), see [18, 13].

In practice, PH machinery is a 3-step process. The first step is the *filtration* step, where one induces a sequence of simplicial complexes from the data. The second step is obtaining the *persistence diagrams*, where PH machinery records the evolution of topological features (birth/death times) in the filtration sequence. The final step is the *vectorization* step where one can convert these records to a feature vector to be used in suitable ML models. For more details on how to apply PH in image analysis, check out the references given in Section 2.2.

Step 1 - Constructing Filtrations: As PH is basically the machinery to keep track of the evolution of topological features in a sequence of simplicial complexes, the most important step is the construction of this sequence. In the case of image analysis, the most common method is to create a nested sequence of binary images (aka cubical complexes). For a given image $\mathcal{X}$ (say $r \times s$ resolution), to create such sequence, one can use grayscale (or other color channels) values γ_{ij} of each pixel $\Delta_{ij} \subset \mathcal{X}$. In particular, for a sequence of grayscale values $(t_1 < t_2 < \cdots < t_N)$, one obtains nested sequence of binary images $\mathcal{X}_1 \subset \mathcal{X}_2 \subset \cdots \subset \mathcal{X}_N$ such that $\mathcal{X}_n = \{\Delta_{ij} \subset \mathcal{X} \mid \gamma_{ij} \leq t_n\}$ (See Figure 2). In other words, we start with a blank $r \times s$ image and start activating (coloring black) pixels when their grayscale value reaches the given threshold. This is called *sublevel filtration* for $\mathcal{X}$ with respect to a given function (grayscale in this case). One can also go in decreasing order to activate the pixels, which is called *superlevel filtration*. Note that if one uses all dimensions for persistent homology, then for cubical persistence, sublevel, and superlevel filtrations have basically the same information by celebrated Alexander duality theorem [28].

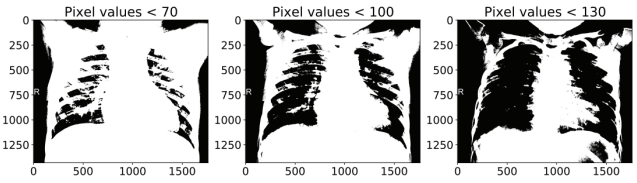


Figure 2: Binary images $\mathcal{X}_{70}, \mathcal{X}_{100}, \mathcal{X}_{130}$ obtained from a chest X-ray for threshold values 70, 100, 130.

Step 2 - Persistence Diagrams The second step in PH process is to obtain persistence diagrams (PD) for the filtration $\mathcal{X}_1 \subset \mathcal{X}_2 \subset \cdots \subset \mathcal{X}_N$, i.e., the sequence of cubical complexes (binary images). PDs are formal summaries of the evolution of topological features in the filtration sequence. PDs are collections of 2-tuples, $\{(b_\sigma, d_\sigma)\}$, marking the birth and death times of the topological features appearing in the filtration. In other words, if a topological feature σ appears for the first time at $\mathcal{X}_{i_0}$, we mark the birth time $b_\sigma = i_0$. Then, if the topological feature σ disappears at $\mathcal{X}_{j_0}$, we mark the death time $d_\sigma = j_0$. i.e., $\mathrm{PD_k}(\mathcal{X}) = \{(b_\sigma, d_\sigma) \mid \sigma \in H_k(\widehat{\mathcal{X}}_i)$ for $b_\sigma \leq i < d_\sigma\}$. Here, $H_k(\mathcal{X}_i)$ represent k^{th} homology group of $\mathcal{X}_i$, representing k-dimensional topological features (k-holes) in cubical complex $\mathcal{X}_i$. By construction, for $2D$ image analysis, only meaningful dimensions to use are $k = 0$ and $k = 1$, i.e., $\mathrm{PD}_0(\mathcal{X})$ and $\mathrm{PD}_1(\mathcal{X})$. For example, 0-dimensional features are connected components and 1-dimensional features are the holes (loops). In our case, if a loop τ first appears at the binary image $\mathcal{X}_3$ and it gets filled in the binary image $\mathcal{X}_7$, we add 2-tuple $(3, 7)$ in the persistence diagram $\mathrm{PD}_1(\mathcal{X})$. Similarly, if a new connected component appears in the binary image $\mathcal{X}_5$ and it merges to the other components in the binary image $\mathcal{X}_8$, we add $(5, 8)$ to $\mathrm{PD}_0(\mathcal{X})$. In Figure 3, we have $\mathrm{PD}_0(\mathcal{X}) = \{(1, \infty), (1, 2), (1, 3), (1, 3), (1, 4), (2, 3)\}$ and $\mathrm{PD}_1(\mathcal{X}) = \{(3, 5), (3, 5), (4, 5)\}$.

This step is pretty standard and there are various software libraries for this task. To obtain PDs for image data with cubical complexes, see [9]. For other types of data and filtrations, see [44].

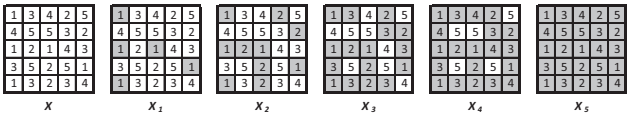


Figure 3: **Sublevel filtration.** The leftmost figure represents an image of 5×5 size with the given pixel values. Then, the sublevel filtration is the sequence of binary images $\mathcal{X}_1 \subset \mathcal{X}_2 \subset \mathcal{X}_3 \subset \mathcal{X}_4 \subset \mathcal{X}_5$.

Step 3 - Vectorization: PDs being a collection of 2-tuples are not very practical to be used with ML tools. Instead, a common way is to convert PD information into a vector or a function, which is called *vectorization* [3]. A common function for this purpose is the *Betti function*, which basically keeps track of the number of "alive" topological features at the given threshold. In particular, the Betti function is a step function with $\beta_0(t_n)$ the count of connected components in the binary image $\mathcal{X}_n$, and $\beta_1(t_n)$ the number of holes (loops) in $\mathcal{X}_n$. In ML applications, Betti functions are usually taken as a vector $\vec{\beta}_k$ of size N with entries $\beta(t_n)$ for $1 \leq n \leq N$, i.e., $\overrightarrow{\beta_k}(\mathcal{X}) = [\beta_k(t_1) \ \ldots \ \beta_k(t_N)]$. e.g., for the image $\mathcal{X}$ in Figure 3, we have $\overrightarrow{\beta_0}(\mathcal{X}) = [5\ 5\ 2\ 1\ 1]$ and $\overrightarrow{\beta_1}(\mathcal{X}) = [0\ 0\ 2\ 3\ 0]$, e.g., $\beta_0(1) = 5$ is the count of components in $\mathcal{X}_1$ and $\beta_1(4) = 3$ is the count of holes (loops) in $\mathcal{X}_4$.

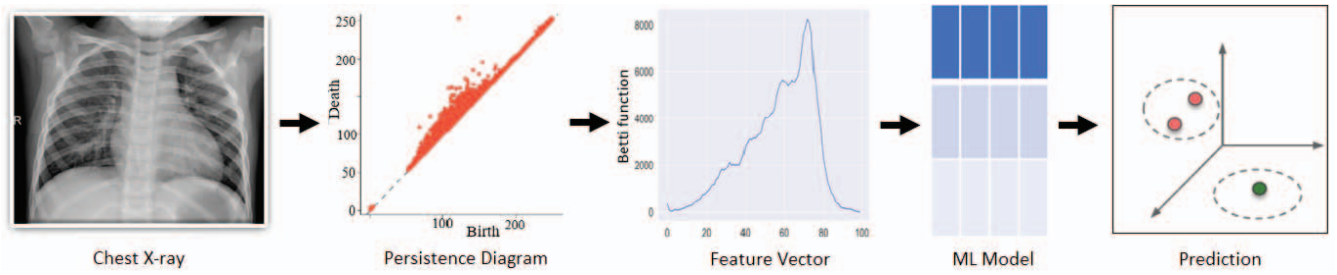


Figure 4: **Flowchart of our model:** For any chest X-ray image, we first get their persistence diagrams by using grayscale values. Then, we obtain our topological feature vectors (Betti functions) out of these persistence diagrams. We feed these vectors to our ML models (RF, XGBoost, etc.) which give highly accurate classification results.

One can consider TDA as a powerful feature extraction method which captures the shape patterns in the image, where the Betti vectors $\vec{\beta}_0(\mathcal{X})$ and $\vec{\beta}_1(\mathcal{X})$ are the corresponding feature vectors. Note that there are various ways to convert PDs into vector in PH (vectorization), e.g, Persistence landscapes, Persistence Images, Silhouettes [3]. Depending on the data type, the choice of vectorization method can be crucial for the performance of the model. In general, the topological features with a short lifespan are considered as *topological noise*. While the other vectorizations try to avoid topological noises, the Betti function takes them into account as well as the dominant features. In chest X-ray images, most topological features have short lifespans, and hence Betti functions work quite well to capture the topological patterns in this case. Furthermore, among these vectorizations, Betti functions are easiest to interpret as being the count of topological features. Because of these reasons, we use Betti functions as vectorizations in this study.

4. Topo-CXR Model for Chest X-ray Screening

In this part, we first describe our model and topological descriptors of chest X-rays. Then, we elaborate on the explainability and interpretability of our model.

4.1. Topological Descriptors for Chest X-rays

In the flowchart (Figure 4), we summarized our Topo-CXR model. Since all chest X-rays (CXR) are grayscale images, it offers a natural filtration method for our persistent homology (PH) approach (Section 3). By using the grayscale pixel values, for any CXR image $\mathcal{X}$, we define sublevel filtration as described in Section 3. While grayscale values varies from 0 (black) to 255 (white), we chose the number of thresholds as $N = 100$ in our filtration step, as further increasing the threshold steps did not increase the performance of our model. In other words, we normalized $[0, 255]$ grayscale interval to $[0, 100]$. After defining the filtration $\mathcal{X}_0 \subset \mathcal{X}_1 \subset \cdots \subset \mathcal{X}_{100}$, we obtain the persistent diagrams $\mathrm{PD}_k(\mathcal{X})$ of each CXR image $\mathcal{X}$ for dimensions $k = 0, 1$ (Section 3). As CXRs are $2D$-images, only $k = 0, 1$ are meaningful dimensions for PH.

After getting persistent diagrams, we convert them into feature vectors as explained in Section 3. In this vectorization step, one can use several choices like Betti functions, Silhouettes, or Persistence Images. Since most of the topological features have short life spans, Betti functions were the natural choice as they give the count of topological features at a given threshold. Furthermore, to keep our model interpretable, we use Betti functions in our models. Hence, we convert $\mathrm{PD}_0(\mathcal{X})$ and $\mathrm{PD}_1(\mathcal{X})$ into corresponding Betti function $\beta_0(\mathcal{X})$ and $\beta_1(\mathcal{X})$ as our feature vectors (Figure 4-Step 2). From Figure 1, one can see that count of topological features at given threshold ranges from 0 to 3500 in $\beta_0(\mathcal{X})$ and 0 to 6000 in $\beta_1(\mathcal{X})$. Note that since our number of thresholds $N = 100$, both $\beta_0(\mathcal{X})$ and $\beta_1(\mathcal{X})$ are 100-dimensional vectors, i.e., $\vec{\beta}_k(\mathcal{X})$ (Section 3). Hence, our topological feature extraction process produce 200 features for any CXR-image $\mathcal{X}$. In Figure 1 and 5, we observe the significant difference between the Betti functions of normal and abnormal CXR images for Pneumonia and Tuberculosis. This figures prove how powerful our topological feature extraction method is.

Then, to use ML tools more effectively, we induce functions (topological summaries) out of these persistence diagrams.

4.2. ML Model

After obtaining our topological feature vectors, the final step is to apply ML tools to these topological fingerprints. To keep our model computationally feasible, we applied tree-based ML methods like Random Forest, XGBoost to our extracted topological features. Since the high number of thresholds ($N = 100$) can bring high collinearity of the output, we used feature selection methods to improve the performance of our model. We give the details of our ML steps in the experiments section (Section 5.2). We note that we did not use any data augmentation and data pre-processing for our model. This makes our model computationally very feasible, and can easily be applied to very large datasets. See Section 6 for further discussion.

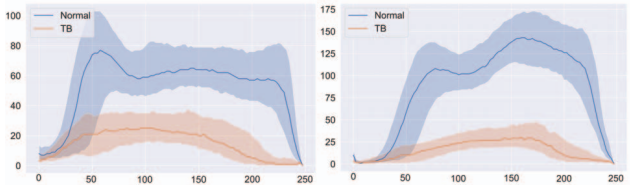
4.3. Explainability and Interpretability

As mentioned in the introduction, one of the main advantages of our model is explainability and interpretability. In Figure 1 and 5, we illustrate the topological patterns created by each class in Pneumonia and Tuberculosis chest X-ray images. In these figures, we give median curves and 40% confidence bands of each class for the corresponding dataset. In supplementary material, we give details of these non-parametric confidence bands and median curves. The distinct topological features for normal and abnormal classes in Figures 1 and 5 explain our model as our topological feature vectors embed CXR images into completely separate clusters in the latent space.

For the interpretability of our model, we need a closer look to our figures. In Figure 1a, we give Betti-0 curves. Recall that in grayscale, the value 0 represents black and 255 represents white. Here, the x-axis represents the grayscale value in [0,255], y-axis represents the count of components. Hence, for a chest X-ray image $\mathcal{X}$, for grayscale value $t \in [0, 255]$, $\beta_0(t)$ represent the number of components in the binary image $\mathcal{X}_t$ (See Figure 2). In Figure 1b, we give Betti-1 curves where the y-axis represents the count of loops. Similarly, for grayscale value $t \in [0, 255]$, $\beta_1(t)$ gives the count of holes/loops in the binary image $\mathcal{X}_t$ (See Figure 2 and 3). For example, in Figure 1a, we observe that median curve for normal class have $\beta_0(100) \sim 1500$ and $\beta_0(180) \sim 3000$. This means, in this dataset, normal chest X-rays have about 1500 components at the grayscale value 100 while they have about 3000 components at the grayscale value 180.

When we interpret Figure 1a in detail, we observe that normal chest X-rays (CXR) develop much more components than pneumonia ones. This interprets as we increase the grayscale value, the binary image (Figure 2) we get for normal CXR becomes highly disconnected, and develops much more components than the pneumonia ones. This means in normal CXRs when you remove lighter parts (grayscale $> t$) from the image, it becomes highly disconnected. However, when you remove lighter parts from the pneumonia CXRs, the image does not become as disconnected as the normal CXRs. The same observation is true for Fig. 1b, too. Normal CXRs develop thousands of holes (loops) in binary images when we increase the grayscale value, pneumonia images develop only a few hundred such holes. These images explicitly show that our topological feature vectors (Betti functions) are very powerful to distinguish these two classes.

In the TB case, our figures again show clear distinction between the topological patterns in normal and abnormal images. For TB-CXR dataset, in Figure 5a and 5b, the number of components and holes is much lower in TB images compared to normal images. In both pneumonia and TB cases, we observe that the induced binary images X_t of the normal class is very spread out and contains much more holes than the abnormal images for most values $t \in [100, 200]$. Note that the number of components and holes are highly different in Ped-Pneumonia dataset and TB-CXR dataset because of the different resolutions of the datasets.



(a) TB vs. Normal (Betti-0) (b) TB vs. Normal (Betti-1)

Figure 5: In the figures above, we give the median curves and 40% confidence bands of our topological feature vectors (Betti functions) for TB-CXR dataset. x-axis represents grayscale values and y-axis represents count of components (Betti-0) or count of loops (Betti-1).

From ML perspective, our feature extraction method is basically an image embedding or fingerprinting approach. For any image, we get 100-dimensional Betti-0, and 100-dimensional Betti-1 vectors (Grayscale values [0,255] normalized to [0,100] in our persistent diagram construction). With these vectors, we embed each image into the latent space $\mathbb{R}^{100}$ (or $\mathbb{R}^{200}$ if both vectors are used). The thinness of the confidence bands shows that each class accumulates in one cluster in the latent space for both feature vectors. In general, for such a classification problem, each class develops several clusters, and ML methods try to distinguish them with different algorithms. In our case, one can consider that the median curves represent the center of the cluster for each class, and the feature vector of each image in that class lands somewhere nearby in the latent (feature) space. The separation between classes and the thinness of the bands prove how powerful the topological features are in these cases.

In Supplementary Material, we give further figures and discussion for our topological feature vectors for TB-Shenzen (Figure 6) and Viral vs. Bacterial Pneumonia classes for Ped-Pneumonia (Figure 7).

5. Experiments

In this section, to evaluate the performance of our Topo-CXR model, we compare it with the current state-of-the-art deep learning models on the publicly available, benchmark datasets for thoracic diseases.

5.1. Datasets

In our experiments for chest X-ray image screening, we used well-known publicly available benchmark datasets. The statistical details of all the dataset we used in this study can be found in Table 1 below.

Table 1: Benchmark datasets for chest X-ray images.

Summary Statistics of Benchmark Datasets					
Dataset	**Image size**	**Total**	**Normal**	**Abnormal**	**Disease**
Ped-Pneumonia [37]	$1914 \times 1628^{*}$	5856	1583	4273	Pneumonia
TB CXR [50]	512×512	4200	3500	700	TB
Shenzhen CXR [33]	3000×3000	662	326	336	TB
CXR-14 [68]	1024×1024	112120	60361	51759	14 Thoracic Dis.

Pediatric Pneumonia (Ped-Pneumonia) CXR dataset [36] is one of the largest publicly available datasets. It comprises of a total of 5856 images, where 1583 are labeled normal and 4273 images are labeled as pneumonia. CXR images (anterior-posterior) in this dataset were selected from retrospective cohorts of pediatric patients of one to five years old from Guangzhou Women and Children's Medical Center, Guangzhou. The resolution of the Ped-Pneumonia images varies, with some images having a minimum resolution of 912×672 pixels and others having a maximum resolution of 2916×2583 pixels.

Shenzhen (CHN) dataset [33] was originally collected in collaboration with Shenzhen No.3 People's Hospital, Guangdong Medical College, Shenzhen, China. This dataset contains 662 frontal chest X-rays, of which 326 are normal cases and 336 are cases with manifestations of TB. In our experiment, we considered all the images in this dataset. All image resolutions are approximately 3000×3000 pixels.

TB-CXR dataset [50] is a publicly available dataset on Kaggle, accessible at the link[1]. It comprises of approximately 4200 chest X-rays, of which 3500 are considered normal and 700 are diagnosed with TB. The dataset is combination of several datasets on TB, namely *NIAID TB dataset* [50], *RSNA CXR dataset* [65], *Belarus CXR dataset* [64], *Shenzhen (CHN) dataset* [33], *Montgomery County (MC) dataset* [33]. All image resolutions are 512×512 pixels.

CXR-14 dataset [68] contains $112,120$ frontal-view X-ray images of 30805 unique patients with each image annotated by [68] with up to 14 different thoracic pathology labels using automatic extraction methods on radiology reports. The 14 thoracic pathology labels are; Infiltration, Effusion, Atelectasis, Nodule, Mass, Pneumothorax, Consolidation, Pleural Thickening, Cardiomegaly, Emphysema, Edema, Fibrosis, Pneumonia, and Hernia. Their corresponding respective number of images are; 19894, 13317, 11559, 6331, 5782, 5302, 4667, 3385, 2776, 2516, 2303, 1686, 1431, 227 and 60361 for no findings [71]. All image resolutions are 1024×1024 pixels.

[1] https://www.kaggle.com/datasets/tawsifurrahman/tuberculosis-tb-chest-xray-dataset

5.2. Experimental Setup

In this part, we give the details of experiments and training of our ML model.

Training:Test Split: Since the majority of datasets (Table 1) does not have a predefined training:test split, many models used their own split. In our accuracy tables below, for each model, we specified their training:test split as it has a significant effect on the performance of the model. Because of this discrepancy between the experimental setups of different methods, we give the basic details of each method in our accuracy tables to facilitate a fair comparison. We used 80:20 splits for all datasets, which is the most common split in all datasets. We used 80% data to build/train the model and our all predictions are based on 20% test (unseen) data.

No Data Augmentation: Note that because of the limited data, all CNN and other deep learning methods have to use serious data augmentation (sometimes 50-100 times) to train their model and avoid overfitting [27]. Our Topo-CXR model are using topological feature vectors, and our feature extraction method is invariant under rotation, flipping and other common data augmentation techniques. Hence, we do not use any type of data augmentation or pre-processing. This makes our model computationally very efficient, and highly robust against small alterations in the image.

ML Model: To increase the performance of our model in terms of accuracy and computational efficiency, we performed parametric tuning and feature selection methods. For feature selection, we used **SelectFromModel** from scikit-learn. We first assign importance to each feature. Then we sorted them in descending order according to threshold parameters. The features are considered unimportant and removed if the corresponding importance of the feature values is below the provided threshold parameter. Random Forest and Extreme Gradient Boosting (XGBoost) ML models are trained on all of the datasets. We used learning_rate = 0.02 and max_depth = 19 for XGBoost. However, after feature selection and fine-tuning models, XGBoost outperformed all the datasets. We obtained the best results with 87 features for Ped-Pneumonia, 67 features for TB-CXR, and 53 features for Shenzhen (CHN) TB datasets. We used Giotto-TDA [62] to obtain persistence diagrams and Betti functions. Our code is available at the link [2].

5.3. Computational Complexity & Runtime:

While PH calculation for high dimensional data is computationally expensive [44], for image data, it is highly efficient. For 2D images, PH has time complexity of $\mathcal{O}(|\mathcal{P}|^r)$ where $r \sim 2.37$ and $|\mathcal{P}|$ is the total number of pixels [42]. In other words, PH computation increases almost quadratic

[2] https://github.com/FaisalAhmed77/Topo-Med

with the resolution. The remaining processes (vectorization, ML) are negligible compared to PH step.

We did all our experiments on a personal laptop with a processor Intel(R), Core(TM) i7-8565U, CPU 1.80GHz, and RAM 16 GB. We mentioned here only one dataset's time complexity. In our experiment, it took only 7443 seconds (2 hours around) to extract Betti-1 features (including PH calculation time) from the TB CXR dataset (4200 images). For the ML part (XGBoost), it took only 1.52 seconds. Runtime for small datasets to extract topological features takes much less time. We used Jupyter notebook as an IDE for writing the code in Python 3.

5.4. Results

In this part, we present the performance of our Topo-CXR model along with state-of-the-art deep learning models for Pneumonia and Tuberculosis(TB) screening on benchmark datasets. In our accuracy tables below, Table 2, 3, 4, 5, we provide the most common performance metrics reported by other models. We used the training data to build the model only. Our all predictions and performance are based on the test (unseen) data. Scikit-Learn [47] library is used for performance metric calculations.

In Tables 2, 3, and 4, the first column gives the total number of images for each dataset used for the corresponding model's experiments. The second column (Train:Test) describes the experimental setup of the model. The third column (# Classes) explains if the task is multilabel or binary. In these tables, we report different performance metrics (AUC, Accuracy, Precision, Recall, etc.) because of the availability of the results in the models compared. In all tables, the best result for each column is given in bold, and the second best result is underlined. For missing data in the table from reference papers, we used "-". We give our results at the bottom row of each table.

Results for Ped-Pneumonia: We give our Pneumonia screening results in Table 2. In Pneumonia screening, we see our model significantly outperforms the state-of-the-art deep learning models. Both Betti-0 and Betti-1 gave very competitive results, but Betti-1 gave the best accuracy 99.7% and AUC 99.95. As Figure 1 suggests, the feature vectors are very distinct for the two classes, and hence such performance for our model is expected. Being computationally very feasible and avoiding data pre-processing, our model can be considered a very promising approach to developing a clinical-decision support method for pneumonia screening from chest X-rays.

Results for TB-CXR Dataset: We give our results for TB diagnosis in Table 3. Our results outperformed again all state-of-the-art deep learning models classifying TB. We got accuracy 99.3% and AUC 99.8 for TB datasets. This superior performance is again not surprising considering the distinct behavior of two classes in Figure 5a and 5b. Since the other papers did not mention their precision/recall scores, we did not include them in the table. Our precision in this dataset is 98.08%, and our recall is 99.29%.

Table 2: Accuracy results for Pneumonia diagnosis on Ped-Pneumonia dataset for binary classification (Pneumonia vs. Normal). Best results are given in bold, and the second best result results are underlined.

Ped-Pneumonia Dataset

Method	Train:Test	Recall	Precision	Accuracy	AUC
xAI [37]	80:20	93.2	90.1	92.8	96.8
mRMR [63]	90:10	96.8	96.9	96.8	96.8
S-CNN [55]	5-fold	94.5	94.3	94.4	94.5
xVGG16 [4]	90:10	89.1	91.3	84.5	87.0
DCNN [49]	92:8	99.0	<u>97.0</u>	<u>98.0</u>	98.0
VGG16 [51]	90:10	99.5	<u>97.0</u>	96.2	99.0
CxNet [71]	77:23	<u>99.6</u>	93.3	96.4	<u>99.3</u>
Topo-CXR	80:20	**99.8**	**99.7**	**99.7**	**99.9**

Table 3: Accuracy results for TB diagnosis on TB-CXR dataset for binary classification (TB vs. Normal).

TB-CXR dataset

Method	# images	Train:Test	Accuracy	AUC
GoogleNet [72]	800	80:20	94.9	-
E-CNN [30]	800	90:10	86.4	-
sCNN [46]	1104	80:20	84.4	92.5
E-CNN [20]	893	70:30	88.8	-
VGG16 [41]	1007	80:20	99.0	98.0
DCNN [50]	7000	80:20	<u>98.6</u>	-
Topo-CXR	4200	80:20	**99.3**	**99.8**

Table 4: Accuracy results for TB diagnosis on Shenzhen (CHN) dataset for binary classification (TB vs. Normal).

Shenzhen (CHN) TB Dataset

Method	# Train:Test	Accuracy	AUC
F-SVM [34]	80:20	84.0	92.5
CNN [32]	70:30	83.7	92.6
sCNN [46]	80:20	84.4	90.0
PT-CNN [40]	5-fold	83.4	91.2
ResNet-BS [52]	90:10	<u>88.8</u>	**95.4**
Topo-CXR	80:20	**89.5**	<u>93.6</u>

Table 5: Classwise AUC results for six thoracic diseases from CXR-14 dataset.

CXR-14 Dataset

Model	Pntx	Eff	Card	Ede	Emph	Mass
DCNN [69]	79.9	75.9	81.0	80.5	83.3	69.3
ResNet50 [5]	84.6	82.8	<u>87.5</u>	84.6	**89.5**	82.1
DenseNet [73]	80.5	80.6	85.6	80.6	84.2	77.7
p-ResNet [39]	<u>87.1</u>	<u>85.9</u>	87.1	88.1	87.0	**83.1**
MobileNet [59]	**88.0**	**87.6**	**88.5**	<u>88.4</u>	<u>89.1</u>	<u>82.6</u>
Topo-CXR	75.8	79.6	80.9	**94.8**	78.5	73.2

Results for Shenzhen TB dataset: We give our results for TB diagnosis in Table 4. Our results are again very competitive with the latest DL models for the TB Shenzhen dataset, where we got accuracy 89.47% and AUC 93.6. For this dataset, we provide the figure for our topological vectors in Figure 6 in the supplementary material. Again, since most papers did not give their precision/recall scores, we did not include them in the table. Our precision in this dataset is 91.07%, and our recall is 84.61%.

Results CXR-14 dataset: To see the performance of our model in other thoracic diseases, we chose the following six additional pathologies out of 14 thoracic diseases: Pneumothorax, Effusion, Cardiomegaly, Edema, Emphysema and Mass. In our experiments, we take a small subset of CXR-14 dataset with Pneumothorax (300), Effusion (301), Cardiomegaly (310), Edema (349), Emphysema (304) Mass (304) and Healthy (474). Even with this small dataset size, and 80:20 training-test split, we get competitive results with SOTA models. We compared our results with current deep learning methods. For this dataset, we used the class-wise AUC results reported in [59, Table 5] where they use 65K images with 80:20 train-test split. Furthermore, our accuracy results are as follows: Pneumothorax (73.55), Effusion (76.77), Cardiomegaly (77.71), Edema (91.52), Emphysema (74.36), and Mass (69.87).

5.5. Discussion

As our figures suggest, our topological descriptors are highly effective in distinguishing between abnormal and normal classes for TB and Pneumonia. Remarkably, even without employing deep learning techniques, our model surpasses state-of-the-art models in benchmark datasets for these diseases (Table 2 and 3). Nevertheless, our experiments reveal that when it comes to other thoracic diseases, our topological feature vectors lag behind CNN models (Table 5). Nonetheless, our results clearly demonstrate the remarkable capability of TDA in generating powerful feature vectors from chest X-rays. Consequently, it is apparent that by seamlessly integrating these topological descriptors with deep learning models, they can significantly contribute to the development of an effective clinical decision support system.

5.6. Limitations

Cubical persistence exhibits invariance under rotation, translation, or flipping, which implies that typical data augmentation methods do not offer significant benefits to our topological model. However, this characteristic bestows a positive outcome by enhancing the robustness of the topological descriptors against noise. The primary limitation of our approach lies in the utilization of mixed-resolution datasets. Specifically, since our Betti functions represent the count of topological features at specific grayscale values, they are susceptible to variations in image resolution. Consequently, any changes in the resolution of chest X-ray (CXR) images directly impact the topological vectors. Therefore, when applying our model to datasets containing images of varying resolutions, it becomes necessary to normalize the images accordingly in order to achieve satisfactory performance.

6. Conclusion

Tuberculosis and Pneumonia are both common pulmonary infectious processes that tremendously impact the global population and healthcare systems with preferentially higher morbidity and mortality in developing countries. Our prediction model provides a reliable screening tool in distinguishing normal and abnormal chest X-ray findings from a relatively small dataset with less complexity and more functionality than deep learning techniques. Further, it helps to identify the subpopulation of patients that would most likely benefit from additional confirmatory testing and subsequent medical therapy. It is meant to be used as an aid to frontline clinicians and radiologists to quickly screen large quantities of patients for tuberculosis and pneumonia, particularly those that do not have access to well staffed radiology departments. Examinations with high pretest probability for abnormality, determined by our model, can undergo a prioritized diagnostic evaluation by a clinician, thereby expediting diagnosis. Although the algorithm struggles with other lung pathologies at this time, its utility for pneumonia and tuberculosis is proven through the course of our experiments. As there is a significant need for an automated chest x-ray screening system to help the clinicians, our unique topological feature vectors can critically help any future ML and DL models to boost their performance and aid them to have more robust results.

Acknowledgement

This work was supported by a grant from National Science Foundation (Grant # DMS-2202584) and by a grant from Simons Foundation (Grant # 579977).

References

[1] Aaron Adcock, Daniel Rubin, and Gunnar Carlsson. Classification of hepatic lesions using the matching metric. *Computer vision and image understanding*, 121:36–42, 2014. 2

[2] H Alaskar, A Hussain, B Almaslukh, T Vaiyapuri, Z Sbai, and Arun Kumar Dubey. Deep learning approaches for automatic localization in medical images. *Computational Intelligence and Neuroscience*, 2022, 2022. 2

[3] Dashti Ali, Aras Asaad, Maria-Jose Jimenez, Vidit Nanda, Eduardo Paluzo-Hidalgo, and Manuel Soriano-Trigueros. A survey of vectorization methods in topological data analysis. *arXiv preprint arXiv:2212.09703*, 2022. 3, 4

[4] Enes Ayan and Halil Murat Ünver. Diagnosis of pneumonia from chest x-ray images using deep learning. In *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, pages 1–5. Ieee, 2019. 7

[5] Ivo M Baltruschat, Hannes Nickisch, Michael Grass, Tobias Knopp, and Axel Saalbach. Comparison of deep learning approaches for multi-label chest x-ray classification. *Scientific reports*, 9(1):1–10, 2019. 7

[6] Paul Bendich, James S Marron, Ezra Miller, Alex Pieloch, and Sean Skwerer. Persistent homology analysis of brain artery trees. *The annals of applied statistics*, 10(1):198, 2016. 2

[7] Eric Berry, Yen-Chi Chen, Jessi Cisewski-Kehe, and Brittany Terese Fasy. Functional summaries of persistence diagrams. *Journal of Applied and Computational Topology*, 4(2):211–262, 2020. 2

[8] Mejbah U Bhuiyan, Christopher C Blyth, Rachel West, Jurissa Lang, Tasmina Rahman, Caitlyn Granland, Camilla de Gier, Meredith L Borland, Ruth B Thornton, Lea-Ann S Kirkham, et al. Combination of clinical symptoms and blood biomarkers can improve discrimination between bacterial or viral community-acquired pneumonia in children. *BMC pulmonary medicine*, 19:1–9, 2019. 12

[9] GUDHI Editorial Board. The gudhi project, 2020. https://gudhi.inria.fr/doc/3.3.0/. 3

[10] Adrian Brady, Risteárd Ó Laoide, Peter McCarthy, and Ronan McDermott. Discrepancy and error in radiology: concepts, causes and consequences. *The Ulster medical journal*, 81(1):3, 2012. 1

[11] Erdi Çallı, Ecem Sogancioglu, Bram van Ginneken, Kicky G van Leeuwen, and Keelin Murphy. Deep learning for chest x-ray analysis: A survey. *Medical Image Analysis*, 72:102125, 2021. 1, 2

[12] Pablo G Cámara, Arnold J Levine, and Raul Rabadan. Inference of ancestral recombination graphs through topological data analysis. *PLoS computational biology*, 12(8):e1005071, 2016. 2

[13] Gunnar Carlsson and Mikael Vejdemo-Johansson. *Topological Data Analysis with Applications*. Cambridge University Press, 2021. 3

[14] F. Chazal and B. Michel. An introduction to topological data analysis: fundamental and practical aspects for data scientists. *Frontiers in Artificial Intelligence*, 4, 2021. 3

[15] Xuxin Chen, Ximin Wang, Ke Zhang, Kar-Ming Fung, Theresa C Thai, Kathleen Moore, Robert S Mannel, Hong Liu, Bin Zheng, and Yuchen Qiu. Recent advances and clinical applications of deep learning in medical image analysis. *Medical Image Analysis*, 79:102444, 2022. 2

[16] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017. 2

[17] Lorin Crawford, Anthea Monod, Andrew X Chen, Sayan Mukherjee, and Raúl Rabadán. Predicting clinical outcomes in glioblastoma: an application of topological and functional data analysis. *Journal of the American Statistical Association*, 115(531):1139–1150, 2020. 2

[18] Tamal Krishna Dey and Yusu Wang. *Computational Topology for Data Analysis*. Cambridge University Press, 2022. 3

[19] Olga Dunaeva, Herbert Edelsbrunner, Anton Lukyanov, Michael Machin, Daria Malkova, Roman Kuvaev, and Sergey Kashin. The classification of endoscopy images with persistent homology. *Pattern Recognition Letters*, 83:13–22, 2016. 2

[20] Lucas Gabriel Coimbra Evalgelista and Elloá B Guedes. Computer-aided tuberculosis detection from chest x-ray images with convolutional neural networks. In *Anais do XV Encontro Nacional de Inteligência Artificial e Computacional*, pages 518–527. SBC, 2018. 7

[21] Kathryn Garside, Robin Henderson, Irina Makarenko, and Cristina Masoller. Topological data analysis of high resolution diabetic retinopathy images. *PloS one*, 14(5):e0217413, 2019. 2

[22] Noah Giansiracusa, Robert Giansiracusa, and Chul Moon. Persistent homology machine learning for fingerprint classification. In *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, pages 1219–1226. IEEE, 2019. 2

[23] Jean Dickinson Gibbons and Subhabrata Chakraborti. *Nonparametric statistical inference*. CRC press, 2014. 12

[24] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 2

[25] Barbara Giunti. Tda applications library, 2022. https://www.zotero.org/groups/2425412/tda-applications/library. 2

[26] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[27] Balla Goutam, Mohammad Farukh Hashmi, Zong Woo Geem, and Neeraj Dhanraj Bokde. A comprehensive review of deep learning strategies in retinal disease diagnosis using fundus images. *IEEE Access*, 2022. 6

[28] Allen Hatcher. *Algebraic Topology*. Cambridge University Press, 2002. 3

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[30] Alfonso Hernández, Ángel Panizo, and David Camacho. An ensemble algorithm based on deep learning for tuberculosis classification. In *International conference on intelligent data engineering and automated learning*, pages 145–154. Springer, 2019. 7

[31] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 2

[32] Sangheum Hwang, Hyo-Eun Kim, Jihoon Jeong, and Hee-Jin Kim. A novel approach for tuberculosis screening based

on deep convolutional neural networks. In *Medical imaging 2016: computer-aided diagnosis*, volume 9785, pages 750–757. SPIE, 2016. 7

[33] Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6):475, 2014. 6

[34] Stefan Jaeger, Alexandros Karargyris, Sema Candemir, Les Folio, Jenifer Siegelman, Fiona Callaghan, Zhiyun Xue, Kannappan Palaniappan, Rahul K Singh, Sameer Antani, et al. Automatic tuberculosis screening using chest radiographs. *IEEE transactions on medical imaging*, 33(2):233–245, 2013. 7

[35] Lida Kanari, Paweł Dłotko, Martina Scolamiero, Ran Levi, Julian Shillcock, Kathryn Hess, and Henry Markram. A topological representation of branching neuronal morphologies. *Neuroinformatics*, 16(1):3–13, 2018. 2

[36] Daniel Kermany, Kang Zhang, Michael Goldbaum, et al. Labeled optical coherence tomography (oct) and chest x-ray images for classification. *Mendeley data*, 2(2):651, 2018. 6

[37] Daniel S Kermany, Michael Goldbaum, Wenjia Cai, Carolina CS Valentim, Huiying Liang, Sally L Baxter, Alex McKeown, Ge Yang, Xiaokang Wu, Fangbing Yan, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018. 6, 7

[38] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. 2

[39] Zhe Li, Chong Wang, Mei Han, Yuan Xue, Wei Wei, Li-Jia Li, and Li Fei-Fei. Thoracic disease identification and localization with limited supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8290–8299, 2018. 7

[40] UK Lopes and João Francisco Valiati. Pre-trained convolutional neural networks as feature extractors for tuberculosis detection. *Computers in biology and medicine*, 89:135–143, 2017. 1, 7

[41] Syeda Shaizadi Meraj, Razali Yaakob, Azreen Azman, SN Rum, Azree Shahrel, Ahmad Nazri, and Nor Fadhlina Zakaria. Detection of pulmonary tuberculosis manifestation in chest x-rays using different convolutional neural network (cnn) models. *Int. J. Eng. Adv. Technol.(IJEAT)*, 9(1):2270–2275, 2019. 7

[42] Nikola Milosavljević, Dmitriy Morozov, and Primoz Skraba. Zigzag persistent homology in matrix multiplication time. In *Proceedings of the twenty-seventh Annual Symposium on Computational Geometry*, pages 216–225, 2011. 6

[43] World Health Organization et al. Global tuberculosis report 2021: supplementary material. 2022. 1

[44] Nina Otter, Mason A Porter, Ulrike Tillmann, Peter Grindrod, and Heather A Harrington. A roadmap for the computation of persistent homology. *EPJ Data Science*, 6:1–38, 2017. 3, 6

[45] Asuka Oyama, Yasuaki Hiraoka, Ippei Obayashi, Yusuke Saikawa, Shigeru Furui, Kenshiro Shiraishi, Shinobu Kumagai, Tatsuya Hayashi, and Jun'ichi Kotoku. Hepatic tumor classification using texture and topology analysis of non-contrast-enhanced three-dimensional t1-weighted mr images with a radiomics approach. *Scientific reports*, 9(1):1–10, 2019. 2

[46] F Pasa, V Golkov, F Pfeiffer, D Cremers, and D Pfeiffer. Efficient deep network architectures for fast chest x-ray tuberculosis screening and visualization. *Scientific reports*, 9(1):1–9, 2019. 7

[47] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011. 7

[48] Talha Qaiser, Yee-Wah Tsang, Daiki Taniyama, Naoya Sakamoto, Kazuaki Nakane, David Epstein, and Nasir Rajpoot. Fast and accurate tumor segmentation of histology images using persistent homology and deep convolutional features. *Medical image analysis*, 55:1–14, 2019. 2

[49] Tawsifur Rahman, Muhammad EH Chowdhury, Amith Khandakar, Khandaker R Islam, Khandaker F Islam, Zaid B Mahbub, Muhammad A Kadir, and Saad Kashem. Transfer learning with deep convolutional neural network (cnn) for pneumonia detection using chest x-ray. *Applied Sciences*, 10(9):3233, 2020. 7

[50] Tawsifur Rahman, Amith Khandakar, Muhammad Abdul Kadir, Khandaker Rejaul Islam, Khandakar F Islam, Rashid Mazhar, Tahir Hamid, Mohammad Tariqul Islam, Saad Kashem, Zaid Bin Mahbub, et al. Reliable tuberculosis detection using chest x-ray with deep learning, segmentation and visualization. *IEEE Access*, 8:191586–191601, 2020. 6, 7

[51] Sivaramakrishnan Rajaraman, Sema Candemir, Incheol Kim, George Thoma, and Sameer Antani. Visualization and interpretation of convolutional neural network predictions in detecting pneumonia in pediatric chest radiographs. *Applied Sciences*, 8(10):1715, 2018. 7

[52] Sivaramakrishnan Rajaraman, Ghada Zamzmi, Les Folio, Philip Alderson, and Sameer Antani. Chest x-ray bone suppression for improving classification of tuberculosis-consistent findings. *Diagnostics*, 11(5):840, 2021. 7

[53] Bastian Rieck, Tristan Yates, Christian Bock, Karsten Borgwardt, Guy Wolf, Nicholas Turk-Browne, and Smita Krishnaswamy. Uncovering the topology of time-varying fmri data using cubical persistence. *Advances in neural information processing systems*, 33:6900–6912, 2020. 2

[54] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[55] Arata Andrade Saraiva, DBS Santos, Nator Junior C Costa, Jose Vigno M Sousa, Nuno M Fonseca Ferreira, Antonio Valente, and Salviano Soares. Models of learning to classify x-ray images for the detection of pneumonia using neural networks. In *Bioimaging*, pages 76–83, 2019. 7

[56] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[57] Yara Skaf and Reinhard Laubenbacher. Topological data analysis in biomedicine: A review. *Journal of Biomedical Informatics*, page 104082, 2022. 2

[58] Primoz Skraba, Maks Ovsjanikov, Frederic Chazal, and Leonidas Guibas. Persistence-based segmentation of deformable shapes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 45–52. IEEE, 2010. 2

[59] Abdelbaki Souid, Nizar Sakli, and Hedi Sakli. Classification and predictions of lung diseases from chest x-rays using mobilenet v2. *Applied Sciences*, 11(6):2751, 2021. 7, 8

[60] Bernadette J Stolz, Tegan Emerson, Satu Nahkuri, Mason A Porter, and Heather A Harrington. Topological data analysis of task-based fmri data from experiments on schizophrenia. *Journal of Physics: Complexity*, 2(3):035006, 2021. 2

[61] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*, 2017. 2

[62] Guillaume Tauzin, Umberto Lupo, Lewis Tunstall, Julian Burella Pérez, Matteo Caorsi, Anibal Medina-Mardones, Alberto Dassatti, and Kathryn Hess. giotto-tda: A topological data analysis toolkit for machine learning and data exploration, 2020. 6

[63] M Toğaçar, B Ergen, Z Cömert, and F Özyurt. A deep feature learning model for pneumonia detection applying a combination of mrmr feature selection and machine learning models. *Irbm*, 41(4):212–222, 2020. 7

[64] https://grantome.com/grant/NIH/AAI12021001-1-0-5. Belarus tb database and tb portal. 6

[65] https://www.kaggle.com/c/rsna-pneumonia-detection-challenge/data. Rsna pneumonia detection challenge. 6

[66] Gaël Varoquaux and Veronika Cheplygina. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ digital medicine*, 5(1):48, 2022. 2

[67] R Virkki, T Juven, H Rikalainen, E Svedström, J Mertsola, and O Ruuskanen. Differentiation of bacterial and viral pneumonia in children. *Thorax*, 57(5):438–441, 2002. 12

[68] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, M Bagheri, and R Summers. Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *IEEE CVPR*, volume 7, 2017. 6

[69] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. 7

[70] Larry Wasserman. Topological data analysis. *Annual Review of Statistics and Its Application*, 5:501–532, 2018. 3

[71] Shuaijing Xu, Hao Wu, and Rongfang Bie. Cxnet-m1: anomaly detection on chest x-rays with image-based deep learning. *IEEE Access*, 7:4466–4477, 2018. 6, 7

[72] Ojasvi Yadav, Kalpdrum Passi, and Chakresh Kumar Jain. Using deep learning to classify x-ray images of potential tuberculosis patients. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2368–2375. IEEE, 2018. 7

[73] Li Yao, Eric Poblenz, Dmitry Dagunts, Ben Covington, Devon Bernard, and Kevin Lyman. Learning to diagnose from scratch by exploiting dependencies among labels. *arXiv preprint arXiv:1710.10501*, 2017. 7