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

Faisal Ahmed, Brighton Nuwagira, Furkan Torlak, and Baris Coskunuzer
The University of Texas at Dallas

Abstract

Chest X-ray analysis is crucial for thoracic disease screening and diagnosis but is time-consuming and costly. Machine learning (ML) has attempted to address this, but faces challenges in feasibility, reliability, and interpretability. This paper introduces a novel feature extraction method using topological data analysis (TDA) for chest X-rays. TDA captures distinct patterns in normal and abnormal images related to pneumonia and tuberculosis. By using cubical persistence, we capture these patterns and convert them into powerful feature vectors, which enhancing interpretability. The resulting model, Topo-CXR, outperforms deep learning models without data augmentation, even on small datasets. Additionally, these topological features can bolster future ML and DL models for improved performance and robustness.

Topo-CXR model Flowchart

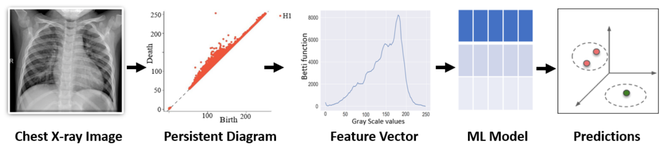
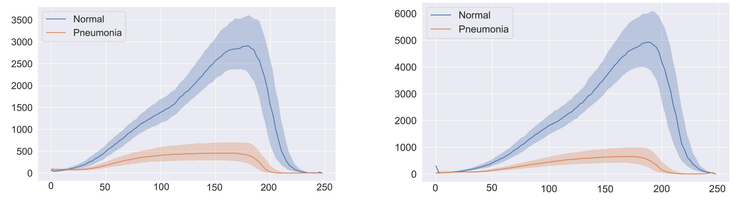


Fig. 1: Flowchart of our model. For any chest X-ray image, we first get their persistence diagrams by using grayscale values. Then, we obtain topological feature vectors (Betti functions) of these persistence diagrams. We feed these functions to our ML models (RF, XGBoost, etc.) which give highly accurate classification (diagnosis) results.



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

Fig 2: The median curves and 40% confidence bands of 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).

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.
- ✓ Analyzing topological patterns in chest X-rays reveals clear distinctions between normal and abnormal images for Pneumonia and TB. We achieve this by utilizing cubical persistence, resulting in powerful feature vectors that capture these unique patterns (refer to Figures 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 , 4, and 5).
- ✓ Unlike many deep learning models, ours operates without the need for data augmentation or preprocessing, delivering exceptional results even with limited datasets. It boasts remarkable computational speed, processing thousands of images in just a few hours.
- ✓ With our powerful topological descriptors, our proposed model is highly explainable and interpretable.

Topo-CXR model steps on Chest X-ray image

Step-1 : Constructing Filtrations

Let X be a grayscale image of $m_x \times m_y$ resolution. Let the square Δ_{uv} represent the pixel with index (u, v) where $u \leq m_x$ and $v \leq m_y$. Let $f : X \to \mathbb{Z}$ represent the filtering function where $f(u, v) \in \{0, 1, \dots, 255\}$ is the grayscale value for the pixel (u, v) . Threshold set is defined by $I = \{\alpha_i\}_{i=1}^{N}$ where $\alpha_1 = 0 < \alpha_2 < \alpha_3 < \dots < \alpha_{N-1} < \alpha_N = 255$. The filtration $\{\widehat{X_i}\}$ is defined by taking the union of pixels $\{\Delta_{uv}\}$ with $f(u, v) \leq \alpha_i$. Then, we obtain the sequence of images $\widehat{X_1} \subset \widehat{X_2} \subset \dots \subset \widehat{X_N}$ (see Fig 3 and 4). This is known as sublevel filtration.

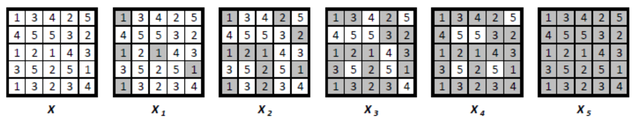


Fig 3: The leftmost figure represents an image of size 5×5 with the given pixel values. Then, the sublevel filtration is the sequence of binary images $x_1 \subset x_2 \subset x_3 \subset x_4 \subset x_5$.

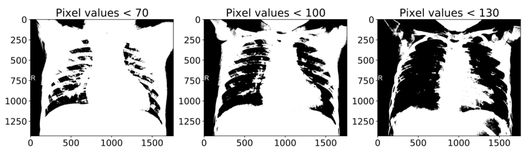


Fig 4: Binary images x_{70}, x_{100}, x_{130} are obtained from a chest X-ray for threshold values 70, 100, 130.

Step-2 : Persistent Diagrams

The second step in PH process is to obtain persistence diagrams (PD) for the filtration $\widehat{X_1} \subset \widehat{X_2} \subset \dots \subset \widehat{X_N}$. PDs are collection of 2-tuples, making the birth and death times of the topological features appearing in the filtration. For instance, if a loop σ appears for the first time at $\widehat{X_{i_0}}$, we mark the birth time, $b_\sigma = i_0$. Then if σ gets filled at $\widehat{X_{j_0}}$, we mark the death time, $d_\sigma = j_0$. In general k-dimensional PD is defined by $PD_k(X) = \{(b_\sigma, d_\sigma) \mid \sigma \in H_k(\widehat{X_i}) \text{ for } b_\sigma \leq i \leq d_\sigma\}$. For 2D image analysis, we will construct $PD_0(X)$ and $PD_1(X)$.

Step-3: Vectorization (Fingerprinting)

Since PH extracts hidden shape patterns from data as persistence diagrams (PD). But PDs being collection of birth and death points in $\mathbb{R}^2$ are not very practical for statistical and ML/DL purposes. One can consider this step as converting PDs into a useful format known as vectorizations. These vectorizations transform PDs into a function or a feature vector form which are much more suitable for ML/DL tools than PDs. Depending on the image dataset, the choice of vectorization is quite important. We use here Betti curves as vectorization for our fingerprints.

- The k^{th} Betti curve $\beta_k : [\alpha_1, \alpha_N] \longrightarrow \mathbb{Z}$ is an integer valued step function where $\beta_k(\alpha_i)$ is the total # barcodes in $PD_k(X)$ containg α_i .

Datasets

Table 1: Summary Statistics of Benchmark Datasets

Dataset	Image size	Total	Normal	Abnormal	Disease
Ped-Pneumonia	$1914 \times 1628^*$	5858	1583	4273	Pneumonia
TB CXR	512×512	4200	3500	700	TB
Shenzhen CXR	3000×3000	662	326	336	TB
CXR-14	1024×1024	112120	60361	51759	14 Thoracic Dis.

EXPERIMENTAL SETUP

- We give the details of our datasets in Table 1. In our experiments, we used most common splits used in the previous works for each dataset. 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.
- 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.
- To increase the performance of our model in terms of accuracy and computational efficiency, we performed parametric tuning and feature selection methods.

Computational Complexity

For image data, PH calculation is highly efficient. For 2D image data, PH computation increases almost quadratic with the resolution. The remaining processes (vectorization, RF) are negligible compared to PH step. We used Giotto-Tda Tauzin et al. (2020) to obtain persistence diagrams, and Betti functions. 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.

Explainability and Interpretability of our results

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

(c) TB-Shenzhen (Betti-0) (d) TB-Shenzhen (Betti-1)

Fig 5: We give the median curves (solid curve) and 40% confidence bands of our topological feature vectors (Betti functions) for TB-CXR dataset and Shenzhen CXR dataset. X-axis represents the grayscale values and Y-axis represents count of components (Betti-0) or counts of loops (Betti-1).

Betti-0 curve gives the count of connected components while Betti-1 gives the count of loops at a given grayscale value. That is $\beta_0(t_0)$ = count of components at grayscale value $T = t_0$ and $\beta_1(t_0)$ = count of loops at grayscale value $T = t_0$.
In Figure 1(a), it becomes evident that the number of components varies significantly between the Pneumonia and Normal classes when considering grayscale values t within the range of [0, 255]. This observed difference highlights a distinct pattern between the Normal and Abnormal classes, providing strong support for the effectiveness of the model.
In Figure 1(b), a similar pattern emerges as the number of loops exhibits significant differences between the Pneumonia and Normal classes for grayscale values t within the range of [0, 255]. This discrepancy underscores a clear distinction between the Normal and Abnormal classes, further reinforcing the robust performance of the model.
Likewise, when examining Figures 5(a) and 5(b), it becomes apparent that there is a distinct pattern between the Normal and TB classes for all grayscale values t within the range of [0, 255]. This pattern strongly supports the exceptional performance of the model in distinguishing between these classes.

RESULTS

Accuracy results on Ped-Pneumonia dataset for binary classification (Pneumonia vs. Normal)

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	97.0	98.0	98.0
VGG16 [51]	90:10	99.5	97.0	96.2	99.0
CxNet [71]	77:23	99.6	93.3	96.4	99.3
Topo-CXR	80:20	**99.8**	**99.7**	**99.7**	**99.9**

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

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	98.6	-
Topo-CXR	4200	80:20	**99.3**	**99.8**

Accuracy results for 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	88.8	**95.4**
Topo-CXR	80:20	**89.5**	93.6

Classwise AUC results for six thoracic diseases from 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	87.5	84.6	**89.5**	82.1
DenseNet [73]	80.5	80.6	85.6	80.6	84.2	77.7
p-ResNet [39]	87.1	85.9	87.1	88.1	87.0	**83.1**
MobileNet [59]	**88.0**	**87.6**	**88.5**	88.4	89.1	82.6
Topo-CXR	75.8	79.6	80.9	**94.8**	78.5	73.2

CONCLUSION

Our prediction model offers a reliable tool for distinguishing normal and abnormal chest X-ray findings, specifically focusing on Tuberculosis and Pneumonia cases. This model, despite being trained on a relatively small dataset, is less complex and more functional than deep learning techniques. It serves as a valuable aid for frontline clinicians and radiologists, particularly in resource-constrained settings, allowing them to efficiently screen large numbers of patients. Patients identified as having a high likelihood of abnormality by our model can undergo prioritized diagnostic evaluation, expediting the diagnosis process. While the algorithm may have limitations with other lung pathologies, it excels in detecting pneumonia and tuberculosis, as demonstrated in our experiments. As there is a pressing need for automated chest X-ray screening systems, our unique topological feature vectors hold the potential to enhance the performance of future machine learning and deep learning models, providing more robust results.