Application of Persistent Homology in the Diagnosis of Breast Cancer from Mammography

Rachel Xing

CS 559 Final Project

Introduction

According to statistics from the American Cancer Society published in 2022 [1], breast cancer rates have been increasing gradually by 0.5% annually in the past four decades, making it the most diagnosed cancer in addition to nonmelanoma skin cancer among women in the United States. However, the survey also highlights a steady decrease in breast cancer mortality since its peak in 1989, with an average annual decline of 1.3% from 2011 to 2022. This improvement in patient outcomes is attributed to advancements in breast cancer diagnosis and treatment, as well as enhanced access to high-quality and early screening. Currently, the primary screening method for breast cancer is mammography, which is an X-ray imaging method to detect any abnormality present in the women's breasts. It can detect small and impalpable tumors and relevant abnormalities such as calcification and architectural distortion that are hard to find by traditional palpation [2]. Accurate interpretation of mammograms requires time and relies heavily on the expertise and experience of pathologists to differentiate between abnormalities and determine whether a tumor is benign or malignant. However, there has been a crisis of pathologist shortages in the United States in the recent decade, with a decrease of 17.53% in the number of pathologists from 2007 to 2017 [3]. This shortage has led to more errors and delays in diagnosis. Therefore, there is a need for a method to streamline the breast cancer diagnosis process, improve the working efficiency of pathologists, and thus optimize patient outcomes. This leads us to consider computer-aided diagnosis as a potential solution.

In recent years, topological data analysis (TDA) has been gaining acceptance and has great potential in a wide range of medical image analysis. Persistent homology, a concept from topology, provides a valuable tool for handling the high dimensionality of medical images and biological data. By using the abstract shapes extracted from concrete images for analysis, TDA allows for direct comparison of images from the same biological system that are collected using different imaging platforms. Additionally, the significant persistent features identified by TDA can offer insights into diagnosis. Recent studies have demonstrated the effectiveness of TDA in differentiating phenotypes of asthma, diagnosing diabetic retinopathy from retinal imaging, and assessing the significance of orthodontic treatment effects [4][5]. Thus, I aim to explore whether TDA in persistent homology can be effectively applied to the diagnosis of breast cancer using mammograms in my project. I hypothesize that the extracted significant persistent features from mammograms can capture diagnostic information about breast cancer because they can handle high-dimensional topological features effectively. Based on this hypothesis, I predict that persistent homology can detect significant differences in mammograms between benign and malignant breast cancer.

Related Work

This project is inspired by a previous study by Gamble & Heo [5] using persistent homology for orthodontic data analysis. In their proposed TDA method, they first calculated persistent homology from 3D landmark-based data for the upper jaw using the Rips filtration method with Euclidean distance to obtain its persistence diagram. They then created a distance matrix from the pairwise distances between persistence diagrams of experimental landmark data. Dimensionality reduction was applied to this high-dimensional distance matrix, converting it to a lower-dimensional matrix suitable for statistical analysis. They found that persistent homology can effectively discern clinical differences in treatment effects among patients in different orthodontic treatment groups and the control group. Some of these clinical differences can be ignored in traditional statistical methods for landmark-based data. In this project, I will adapt this TDA method for analyzing 2D mammograms.

The technique relevant to this TDA method is Giotto-TDA, a Python toolkit specialized for TDA based on scikit-learn, proposed by Tauzin et al.

in 2020 [6]. This toolkit offers more specific and comprehensive calculators and filters for TDA compared to ParaView/TTK. First, it provides more interactive persistence diagrams that can differentiate persistent features with different Betti dimensions, allowing for the assessment of the significance of various persistent features in TDA. It also includes functions for Wasserstein distance and pairwise distance, enabling the computation of distance matrices, which ParaView/TTK and other TDA toolkits like scikit-TDA lack. However, scikit-learn [7] is still used in this study because it provides a useful tool, multidimensional scaling (MDS), for dimensionality reduction. This is a crucial part of our method to transform high-dimensional pairwise distance matrices into a low-dimensional form, allowing for further statistical analysis.

Methods

Dataset and Preprocessing

The dataset used in this study consists of 40 mammograms randomly selected from CBIS-DDSM (Curated Breast Imaging Subset of DDSM). All mammograms are of the right breast from the craniocaudal (RCC) view, showing both of its medial and external lateral parts. The dataset includes 20 mammograms diagnosed with benign breast cancer and 20 mammograms diagnosed with malignant breast cancer. CBIS-DDSM, published by Lee et al. in 2017 [8], is a public dataset of mammograms that includes both MLO (mediolateral oblique) and CC (craniocaudal) views. It contains benign and malignant cases formatted for modern computer vision datasets, suitable for data analysis and machine learning. It is a standardized subset of DDSM (Digital Database for Screening Mammography), published by the University of South Florida (available online at). In addition to benign and malignant cases, DDSM also contains negative cases, which I originally planned to include in this study. However, this dataset requires special permission to access, and the mammograms are not standardized.

For analyzing the 2D images, all topological computations are performed on the matrix of their pixel intensities. However, most sampled mammograms have sizes of 2000×4000 or larger. This means each matrix includes 8,000,000 pixels, which significantly slows down the computation due to the large data size. Therefore, to ensure the efficiency of the computation process, I resized all images to a size of 144×228 (Figure 1) and then transformed them into 3D point clouds. These point clouds can preserve the original image dimensions, with each element corresponding to a pixel's grayscale intensity, the third dimension of the point cloud representation. This mathematical representation of 2D images is more manageable and well-suited for processing by TDA algorithms.

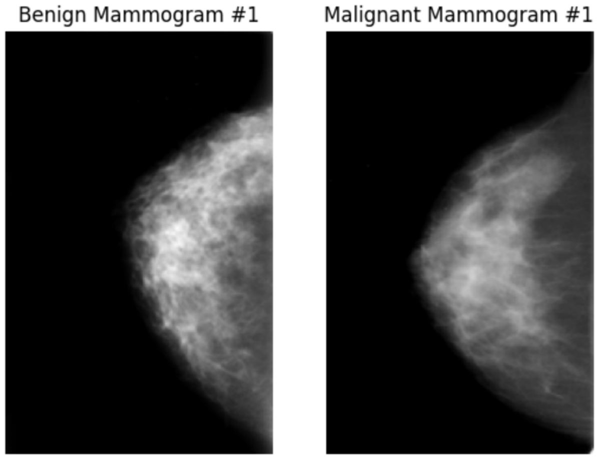

Figure 1: *Resized RCC mammograms diagnosed with benign (left) and malignant (right) breast cancer.*

Persistent Homology Computation

This project uses the Rips filtration using Euclidean distance to compute the persistent diagram for each mammogram in the dataset, which is the same method used in the study by Gamble & Heo [5]. In the Rips filtration, all points in the point cloud are introduced into the topological space at time t = 0, and edges between two points form when their Euclidean distances equal a specific time t. The Euclidean distance between point i and point j in the point cloud is calculated as equation 1:

$$\mathrm{ild}(i,j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \quad (1)$$

where x and y represent the locations of points on the two axes of 2D coordinates, while z represents the pixel intensity of each point in the corresponding mammogram.

As time t progresses, triangles form when all three of their edges have been added, and three-dimensional tetrahedra join when all of their faces are included. Throughout this process, we can observe the birth and death of persistent features. The difference in times between death and birth represents the persistence of these features in the topological space. Since the point cloud originates from 2D images, we specifically focus on β_0 and β_1 persistent features, corresponding to connected components and 1-D holes, respectively [9]. The gtda.homology.VietorisRipsPersistence filter performs this computation, producing a matrix of persistent features with their calculated persistence corresponding to the nth Betti dimension.

Distance Matrix Computation

Then, all persistent diagrams of the 40 mammograms are combined into a single matrix, along with two separate matrices for the extracted diagrams of β_0 and β_1 persistent features. Next, we compute the pairwise distance for each pair of persistent features in each matrix using the Wasserstein distance, which measures the similarity between each pair of diagrams. This approach also follows the methodology in Gamble & Heo's study [5]. The formula for the Wasserstein distance between diagrams a and b is defined as:

$$\gamma_l(x) : \mathrm{Dgm}_l(a) \to \mathrm{Dgm}_l(b) \tag{2}$$

$$W_p(a,b) = [\sum_l \inf_{\gamma_l} \sum_{x \in \mathrm{Dgm}_l(a)} ||x - \underset{\inf}{\overset{p}{\gamma_l}}(x)||]^{1/p} \tag{3}$$

where l refers to the Betti dimension, p refers to the order of matching between two diagrams, and γ is the optimal matching. Higher values of p make the distance more sensitive to larger differences between the diagrams. The computation is performed using the gtda.diagrams.PairwiseDistance function with the metric set to "wasserstein" and the default value of p = 2. The output is a symmetric matrix with dimensions 40×40, where each element represents the Wasserstein distance between a pair of diagrams.

Dimensionality Reduction

It is hard to perform data analysis directly on the high-dimensional 40×40 distance matrices for both benign and malignant mammograms. Therefore, dimensionality reduction is necessary to convert these matrices into lower-dimensional representations. The method used here is Multidimensional Scaling (MDS), as mentioned in the study by Gamble & Heo [5]. Given a high-dimensional matrix of pairwise distances between N objects, MDS projects these objects into a lower K-dimensional space while retaining the original pairwise distances as closely as possible in the new K coordinates [10]. To determine the appropriate number of K dimensions to retain, we first perform Principal Component Analysis (PCA). PCA involves computing the covariance matrix of the distance matrix and then calculating the eigenvalues, which represent the amount of variance explained by each principal component (from 1 to N) [11]. The number of principal components (i.e., dimensions) to retain is decided based on the visualization of cumulative explained variance with increasing dimensions using scree plots. PCA is performed using the sklearn.decomposition.PCA function to identify the number of dimensions to retain. Then, the sklearn.manifold.MDS function is used for multidimensional scaling, and the reduced distance matrix will be an N×40 matrix of MDS coordinates for further data analysis.

Data Analysis

The data analysis on the reduced distance matrix involves using MANOVA (Multivariate Analysis of Variance) to test if the multiple MDS coordinates reduced from the distance matrix computed from persistent homology on mammograms can explain the different types of breast cancer (benign or malignant). This is performed using the statsmodels.multivariate.manova.MANOVA function, with a significance level of 0.05 for this study.

Results and Analysis

Persistent Homology on Each Mammogram

Figure 2 shows the computed persistent homology for benign and malignant mammograms. We can observe that β_0 and β_1 are two persistent Betti dimensions present in the 2D mammograms. However, most features in all persistent diagrams are β_0 features, with only a few β_1 features persisting in the mammograms. In some persistent diagrams, such as that of benign mammogram #5, there is a lack of β_1 features. That may imply that the β_1 feature might lack significance in classifying benign and malignant breast cancer because it does not persist across all mammograms.

Multidimensional Scaling

All scree plots for the three distance matrices show that about 90% of the variance within the distance matrix can be explained by using just two dimensions (Figure 3). Thus, two dimensions are

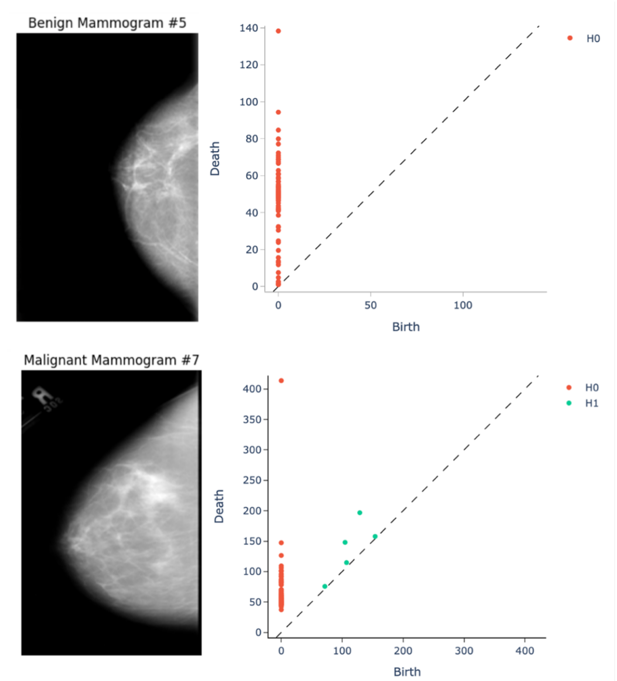

Figure 2: *Benign mammogram #5 (above) and malignant mammogram #7 (below) with their corresponding persistent diagrams, including highlighted β_0 and β_1 features.*

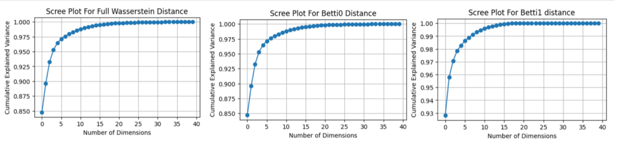

Figure 3: *Scree plots for full Wasserstein (from full persistent diagrams), β_0, and β_1 distance matrices. The x-axis indicates the number of dimensions, and the y-axis indicates the cumulative variance explained.*

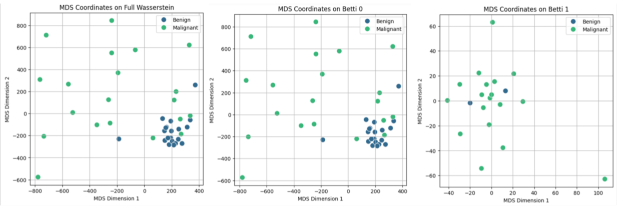

Figure 4: *Scatterplot for the MDS coordinates on the retained two dimensions from full Wasserstein (from full persistent diagrams), β_0, and β_1 distance matrices, labeled by experimental groups.*

Betti Dimensions	F-value	MANOVA p-value
Full	17.2751	<0.0001
β_0	17.3395	<0.0001
β_1	0.2776	0.7592

Table 1: *Results of MANOVAs on the full dimensions-, β_0-, and β_1-based MDS coordinate between the benign and malignant groups. The p-values are derived from Wilks' Lambda, with a significance level set at 0.05.*

chosen to be retained during multidimensional scaling. Then, we use scatterplots to visualize the MDS coordinates reduced from the high-dimensional distance matrix (Figure 4).

From the scatterplot of the first two MDS dimensions in Figure 4, we can find that the differences between the groups are evident in the reduced full distance matrix and β_0 distance matrix. This is demonstrated by the blue points (benign breast cancer) clustering together, while the green points (malignant breast cancer) are distributed in the remaining areas of the MDS coordinates. However, such differences are not discernible in the reduced β_1 distance matrix, implying that β_1 might not provide significant diagnostic information for breast cancer.

Data Analysis

The MANOVA results in Table 1 show that both full persistent homology and β_0 persistent features can effectively explain the diagnostic differences between benign and malignant breast cancer in mammograms (p-value < 0.0001). In contrast, β_1 persistent features cannot achieve this, with a large p-value greater than 0.05. This confirms our hypothesis and previous findings that β_1 persistent features might lack the ability to provide useful topological information for breast cancer diagnosis.

Conclusion and Future Work

In summary, persistent homology can effectively capture significant topological features in mammograms that can help differentiate between benign and malignant breast cancer. Both full persistent homology and β_0 persistent features can capture these diagnostical details. However, β_1 persistent features are less effective in this context, likely because tumors or microcalcifications, the two major indicators of breast cancer, are dense components rather than holes, making it less informative.

This project presents preliminary results showing the potential of persistent homology in computer-aided breast cancer diagnosis. However, further work is needed to ensure its practicability and reliability. First, we need to verify that the capability of persistent homology on breast cancer diagnosis can extend to other views of mammograms and the left breast. Besides, this project only uses benign and malignant cases, whereas the BI-RADS final assessment categorizes breast cancer into six levels: negative, benign, probably benign, suspicious abnormality, highly suggestive of malignancy, and known malignancy [2]. To align with the actual complexity of breast cancer diagnosis, we need to further validate the refined classification of breast cancer using persistent homology. We should also explore the utility of persistent homology in assessing other diagnostic indicators, such as abnormality type and cancer grade. Moreover, a recent study by Vandaele et al. [12] has demonstrated that topological features from persistent homology can aid in diagnosis and improve radiomic-based histology predictions for lung tumors. Similarly, we can investigate how persistent homology can contribute to predicting or monitoring cancer progression in mammograms, which may potentially offer valuable insights for enhancing treatment strategies for breast cancer.

References

[1] A. N. Giaquinto, H. Sung, K. D. Miller, *et al.*, "Breast cancer statistics, 2022," *CA Cancer J. Clin.*, vol. 72, no. 6, pp. 524–541, 2022. DOI: 10.3322/caac.21754.

[2] M. K. Shetty, "Screening for breast cancer with mammography: Current status and an overview," *Indian J. Surg. Oncol.*, vol. 1, no. 3, pp. 218–223, Sep. 2010. DOI: 10.1007/s13193-010-0014-x.

[3] W. Y. Naritoku, M. A. Furlong, B. Knollman-Ritschel, and K. L. Kaul, "Enhancing the pipeline of pathologists in the united states," *Acad. Pathol.*, vol. 8, 2021. DOI: 10.1177/23742895211041725.

[4] Y. Singh, C. M. Farrelly, Q. A. Hathaway, *et al.*, "Topological data analysis in medical imaging: Current state of the art," *Insights Imaging*, vol. 14, no. 1, 2023, Art no. 58. DOI: 10.1186/s13244-023-01413-w.

[5] J. Gamble and G. Heo, "Exploring uses of persistent homology for statistical analysis of landmark-based shape data," *J. Multivar. Anal.*, vol. 101, no. 9, pp. 2184–2199, 2010. DOI: 10.1016/j.jmva.2010.04.016.

[6] G. Tauzin, U. Lupo, L. Tunstall, *et al.*, *Giotto-tda: A topological data analysis toolkit for machine learning and data exploration*, 2021. arXiv: 2004.02551 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2004.02551.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, *Scikit-learn: Machine learning in python*, 2018. arXiv: 1201.0490 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1201.0490.

[8] R. S. Lee, F. Gimenez, A. Hoogi, K. K. Miyake, M. Gorovoy, and D. L. Rubin, "A curated mammography data set for use in computer-aided detection and diagnosis research," *Sci. Data*, vol. 4, no. 1, 2017, Art no. 170177. DOI: 10.1038/sdata.2017.177.

[9] Y. Dabaghian, F. Mémoli, L. Frank, and G. Carlsson, "A topological paradigm for hippocampal spatial map formation using persistent homology," *PLoS Comput. Biol.*, vol. 8, no. 8, pp. 1–14, Aug. 2012. DOI: 10.1371/journal.pcbi.1002581.

[10] C.-C. M. Yeh, H. Van Herle, and E. Keogh, "Matrix profile iii: The matrix profile allows visualization of salient subsequences in massive time series," in *2016 IEEE 16th International Conference on Data Mining (ICDM)*, 2016, pp. 579–588. DOI: 10.1109/ICDM.2016.0069. eprint: 2016.0069.

[11] I. T. Jolliffe and J. Cadima, "Principal component analysis: A review and recent developments," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 374, no. 2065, 2016. DOI: 10.1098/rsta.2015.0202.

[12] R. Vandaele, P. Mukherjee, H. M. Selby, R. P. Shah, and O. Gevaert, "Topological data analysis of thoracic radiographic images shows improved radiomics-based lung tumor histology prediction," *Patterns*, vol. 4, no. 1, 2023, Art no. 100657. DOI: 10.1016/j.patter.2022.100657.