mathematics

MDPI

Article

Persistent Homology for Breast Tumor Classification Using Mammogram Scans

Aras Asaad [1,*], Dashti Ali [2], Taban Majeed [3] and Rasber Rashid [3]

[1] School of Computing, The University of Buckingham, Buckingham MK18 1EG, UK
[2] Independent Researcher, North York, ON M2R 1G4, Canada
[3] Department of Computer Science and IT, Salahaddin University, Erbil 44001, Iraq
* Correspondence: aras.asaad@buckingham.ac.uk

Abstract: An important tool in the field of topological data analysis is persistent homology (PH), which is used to encode abstract representations of the homology of data at different resolutions in the form of persistence barcode (PB). Normally, one will obtain one PB from a digital image when using a sublevel-set filtration method. In this work, we built more than one PB representation of a single image based on a landmark selection method, known as local binary patterns (LBP), which encode different types of local texture from a digital image. Starting from the top-left corner of any 3-by-3 patch selected from an input image, the LBP process starts by subtracting the central pixel value from its eight neighboring pixel values. Then, each cell is assigned with 1 if the subtraction outcome is positive, and 0 otherwise, to obtain an 8-bit binary representation. This process will identify a set of landmark pixels to represent 0-simplices and use Vietoris–Rips filtration to obtain its corresponding PB. Using LBP, we can construct up to 56 PBs from a single image if we restrict to only using the binary codes that have two circular transitions between 1 and 0. The information within these 56 PBs contain detailed local and global topological and geometrical information, which can be used to design effective machine learning models. We used four different PB vectorizations, namely, persistence landscapes, persistence images, Betti curves (barcode binning), and PB statistics. We tested the effectiveness of the proposed landmark-based PH on two publicly available breast abnormality detection datasets using mammogram scans. The sensitivity and specificity of the landmark-based PH obtained was over 90% and 85%, respectively, in both datasets for the detection of abnormal breast scans. Finally, the experimental results provide new insights on using different PB vectorizations with sublevel set filtrations and landmark-based Vietoris–Rips filtration from digital mammogram scans.

Keywords: topological data analysis; persistent homology; breast mammogram; persistence diagram vectorization; medical imaging; local binary patterns

MSC: 55N31

check for updates

Citation: Asaad, A.; Ali, D.; Majeed, T.; Rashid, R. Persistent Homology for Breast Tumor Classification Using Mammogram Scans. *Mathematics* **2022**, *10*, 4039. https://doi.org/10.3390/math10214039

Academic Editors: Rocio Gonzalez Diaz and Matthias Zeppelzauer

Received: 30 September 2022
Accepted: 27 October 2022
Published: 31 October 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Topological data analysis (TDA) is a collection of methods from algebraic topology and geometry to build and extract topological features from data. Persistent homology (PH), the main tool of TDA, extracts topological summaries from data in the form of connected components, loops, and cavities using a process known as filtration, which relies on a nested sequence of simplicial complexes that capture the birth and death of those topological invariants [1]. A collection of births and deaths of these topological features are then represented as points in persistence diagram(s) (PD) or equivalently as bars in persistence barcode(s) (PB). Topological structures represented as PDs are stable with respect to small perturbations to the input data when the bottleneck or Wasserstein distance is used to compare PDs [2]. Although mostly used when the input data have the form of a

point cloud, PH can also be used when the input data to TDA pipeline are images where they have a grid structure. We demonstrate that one can construct Vietoris–Rips filtration from digital images based on pixel landmark locations that convey different types of local textural information. In this paper, we aimed at harnessing the power of PH to differentiate benign breast tumors from their malignant counterparts using breast mammogram. A mammogram scan is a special type of X-ray imaging that involves exposing breast tissues to a small amount of radiation to obtain an inside picture of the breast details for the purpose of abnormality/mass detection and classification.

Female breast cancer is among the four leading types of cancer in women worldwide. The World Health Organization (WHO) and its cancer research agencies such as the American Cancer Society and International Agency for Cancer Research reported 19.3 million new cases of cancer in 2020 with 10 million deaths and estimated that this number could be increased to 28.4 million new cases by 2040 [3]. Mammogram scans have a number of advantages to detect early signs of breast cancer in women, among them being their wide deployment in hospitals, their ease of storage, less time to examine by radiologists, and low cost. A number of difficulties face radiologists in properly examining mammograms such as low resolution, size of the lesion within the breast tissue, location of the lesion, and dense breast tissue in young patients. Therefore, designing sophisticated computer aided diagnostics (CAD) to assist radiologists in making their final decision is a necessity.

The main contribution of this paper can be summarized as follows. (1) Constructing 56 persistence diagrams from a single mammogram whereby each PD is constructed based on a set of automatically extracted mammogram pixel locations that convey different type of textural information. (2) The space of persistence barcodes featurized using four different methods, namely binning, barcode-statistics, persistence images, and persistence landscapes to measure the true performance of the proposed approach.

2. Methods

To build PH from digital mammograms, we relied on pixel-based landmarks that correspond to abnormality in textures. We derived our approach from a texture descriptor method known as local binary patterns (LBPs) introduced more than two decades ago in [4]. Abnormality is expected to distort the local texture and structure in mammogram scans. Using LBP, we encoded this change in the local texture and structure of mammograms to ensemble a set of point clouds as input to the PH pipeline. This method provides a rich source of persistence topological features for machine learning. Next, we describe our proposed landmark selection procedure and PH construction.

2.1. Image Patch Local Binary Patterns (IP-LBPs)

Since 1996, LBP has been used successfully in many pattern recognition applications and different versions of LBP have been proposed and investigated with considerable success [5–7]. For any grayscale image I, LBP constructs a new grayscale image $\bar{I}$ by encoding each pixel $p \in I$ with 8-bit binary representation determined by comparing the central pixel with that of its eight neighbors in a 3-by-3 image-patch, surrounding it in a clockwise manner. Starting from the top-left corner of any 3-by-3 patch, the LBP process starts by subtracting the central pixel value from its eight neighboring pixel values. Then, each cell is assigned with 1 if the subtraction outcome is positive, and 0 otherwise (see Figure 1 for illustration). This process results in an 8-bit binary code that can then be converted back to decimal values representing the central pixel (x_c, y_c) using the following equation:

$$LBP(x_c, y_c) = \sum_{i=1}^{i=8} f\,(p_i - p_c) \times 2^i \tag{1}$$

where p_c is the central pixel value; p_i is the neighboring gray-value pixels; and the function $f(x)$ is defined as follows:

$$f(x) = \begin{cases} 1 & if\ x \geq 0 \\ 0 & if\ x < 0 \end{cases} \tag{2}$$

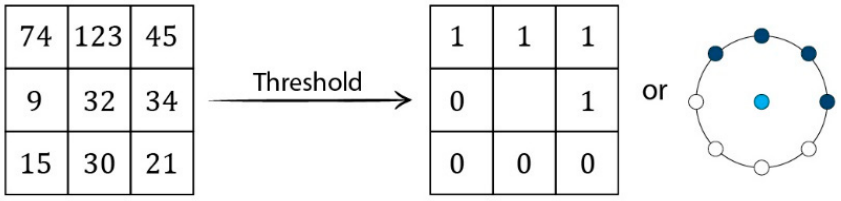

Figure 1. The LBP process where 1's in the binary code is represented by the bold points on the circle.

In total, there are 256 binary codes that one will obtain for any 3-by-3 image patch following the LBP procedure. In [8], Ojala et al. demonstrated that only 58 binary codes out of the 256 were enough to represent 90% of textures in natural images. The 58 binary codes are known as uniform LBP (ULBP) and they experimentally demonstrated that the histogram of ULBP codes can be used as a discriminating feature for computer vision applications [7,9]. ULBP codes encode local texture features such as edges, corners, spots, lines, and flat regions in an image and their binary codes have either 0 or 2 circular transitions from 0 to 1 or from 1 to 0. There are 56 ULBP codes that have two circular transitions and only two ULBP codes with 0 circular transitions in their 8-bit binary representation. 00000000 and 11111111 are the two ULBP codes with 0 transitions. Examples of ULBP codes with two circular transitions are 11000000 and 00111100, whereas a binary code such as 10101010 is not a ULBP because there are more than two circular transitions from 1 to 0 or vice versa. We can group the 56 ULBP codes according to the number of 1's in their binary representation to form a 7-group geometry G_λ for $\lambda = 1, 2, \ldots, 7$ where λ refers to the number of 1's in each geometry. Furthermore, each G_λ consists of eight binary codes that can be obtained from each other by a circular rotation (see Figure 2). Starting from the top-left corner of the mammograms, we scanned the entire input image by selecting the central pixel value of 3-by-3 patches as landmarks if its binary representation satisfied one of the geometrical circles in Figure 2.

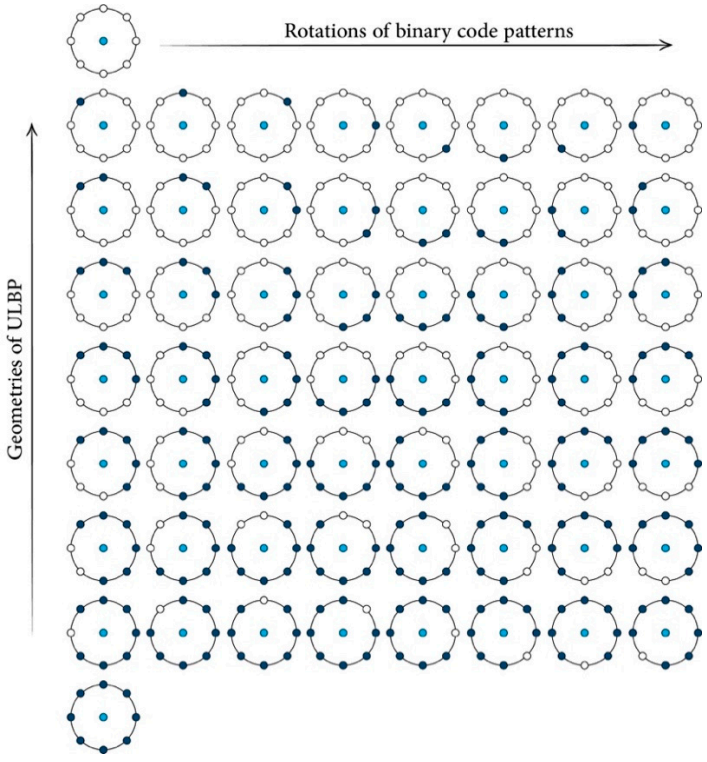

Figure 2. Geometric representation of the ULBP method.

We can select one or more of these geometries to select landmarks from digital mammograms to construct PD. Different rows correspond to different types of texture. For example, the first and the last row correspond to flat and spot texture, row 5 (4 ones in the binary code) corresponds to edges and row 6 corresponds to corners.

For example, candidate landmarks are central pixel positions of the first rotation of $G_1(R_1)$ if a 3-by-3 patch's binary code is 00000001 where R_ξ refers to rotations of specific G_λ for $\xi = 1, 2, \ldots, 8$. We followed the same strategy to select a set of pixel value locations for each of the 56 ULBP geometries depicted in Figure 2.

The first two stages of Figure 3 show an example of a set of landmark pixel locations extracted from a normal and abnormal mammogram and their corresponding PDs. After selecting a set of mammogram pixel landmarks, we generated a Euclidean distance matrix D from these pixel value locations, which will then be used as input for the PH generation pipeline.

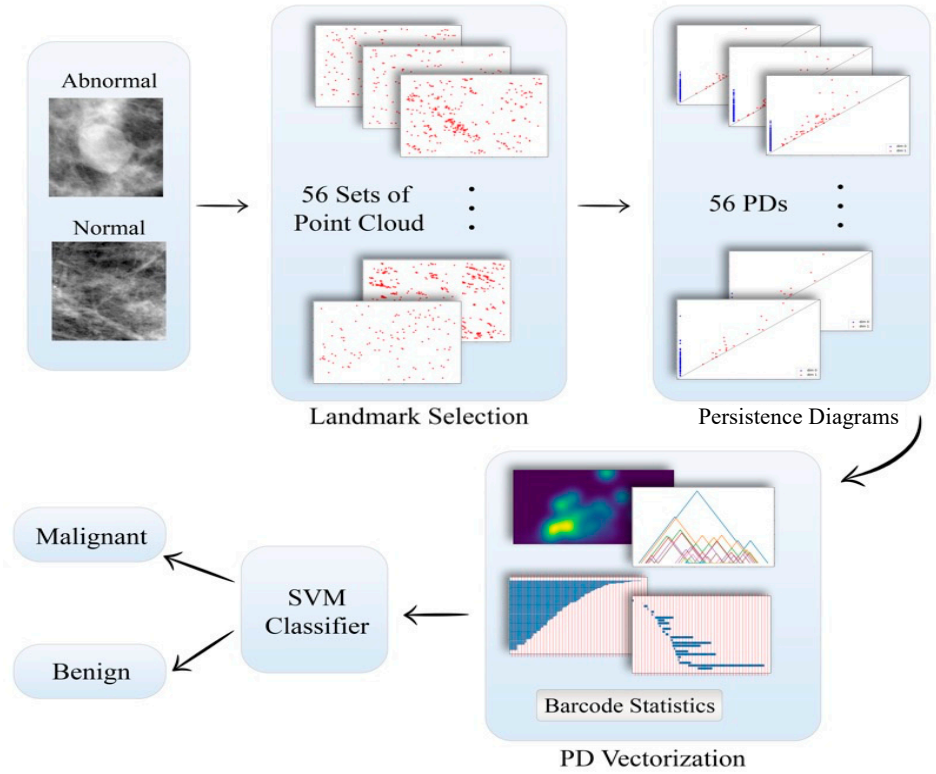

Figure 3. Landmark based PH construction and classification pipeline.

2.2. Persistent Homology of Digital Images

In this section, we introduce Vietoris–Rips simplicial complexes and cubical complexes as two persistent homology approaches to build topological features from breast mammograms.

2.2.1. Vietoris–Rips Complexes Based on Image Pixel Landmarks

In order to build topological features from data (point cloud or image), the PH relies on mathematical objects known as simplices, which are building blocks of higher dimensional objects in space known as the simplicial complex. In this work, we constructed Vietoris–Rips (VR) simplicial complexes using the pixel-value locations obtained from the ULBP method. For a set $\mathcal{L}$ of pixel landmarks in $\mathbb{R}^2$, its VR with parameter ϵ, denoted as $VR(\mathcal{L}, \epsilon)$, is the simplicial complex where $\{l_0, l_1, l_2, \ldots, l_\eta\}$ is its vertex set, which spans a η-simplex if the Euclidean distance between any two landmark locations is less than or equal to the chosen value of ϵ (i.e., $d(l_i, l_j) \leq \epsilon \; \forall \; 0 \leq i, j \leq \eta$). As we increase the value of ϵ, so does the VR of the pixel locations. This process results in producing a nested sequence of VR

simplicial complexes known as filtration. In other words, $VR(\mathcal{L}, \epsilon_1) \subseteq VR(\mathcal{L}, \epsilon_2)$ if $\epsilon_1 \leq \epsilon_2$. Homological features born and vanished during the filtration process are then stored as points in PD.

We direct the interested reader to see [1,10,11] for more mathematical details on VR construction from a point cloud, PH introduction, and mathematical backgrounds, respectively.

2.2.2. Cubical Complexes of Digital Images

The cubical analogue of a (VR) simplicial complex is a cubical complex in which the role of simplices is played by cubes of different dimensions, as shown in Figure 4. A finite cubical complex in R^d is a union of cubes aligned on the grid $\mathbb{Z}^d$, satisfying some conditions similar to the simplicial complex case [12]. A d-dimensional image is a map $\eta : I \subseteq \mathbb{Z}^d \to \mathbb{R}$. An element $v \in I$ is called the voxel, or a pixel when $d = 2$, and $\eta(v)$ is called its greyscale value. There are several ways to represent digital images as cubical complexes, but the greyscale image comes with a natural filtration and was hence adopted here. Voxels are represented by vertices and cubes are built between these vertices. We represent voxels by d-cubes and all of its adjacent lower dimensional cubes are added. Next, we obtained a function on the resulting cubical complex $\mathbb{K}$ by extending the values of voxels to all of the cubes $\sigma \in \mathbb{K}$ as follows:

$$\eta'(\sigma) := \min_{\sigma \text{ face of } \tau} \eta(\tau)$$

Assume $\mathbb{K}$ to be the resulting cubical complex built on the greyscale image I. Let

$$\mathbb{K}_i := \{\sigma \in \mathbb{K} \,|\eta'(\sigma) \leq i\}$$

be the i-th sublevel set of $\mathbb{K}$. The set $\{\mathbb{K}_i\}_{i \in Im(I)}$ defines a filtration of cubical complexes indexed by the value of greyscale function η.

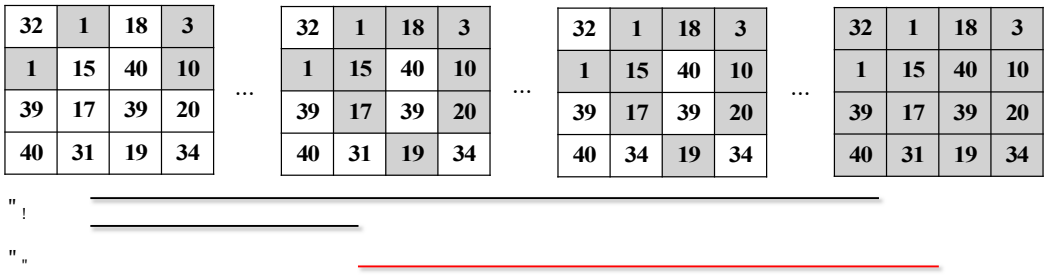

Figure 4. A greyscale image patch and its corresponding cubical complex filtration and persistent barcode representation in dimension zero and one. B_0 and B_1 represent Betti numbers in dimension zero and one, respectively.

2.3. Persistence Diagram Vectorization

Topological features summarized by PDs are not amenable to many machine learning and statistical tasks; for instance, PD's Fréchet mean is not unique [13]. Hence, many vectorization approaches have been proposed to transform the data in PDs to resolve this issue and be able to apply machine learning methods. We used four methods to vectorize the topological features in PD: persistence images [14], persistence landscapes [15], binning [16], and barcode statistics. Next, we briefly describe each of these vectorization approaches.

Persistence landscape (PL). PL is one of the early vectorization methods proposed to map PDs into a stable and invertible function space using a family of piecewise linear functions $\{\Psi_k : \mathbb{R} \to \mathbb{R}\}_{k \in \mathbb{Z}}$ so that $\Psi_k(\tau) = \sup\{m \geq 0 \mid \alpha^{\tau-m,\, \tau+m} \geq k\}$, where $\alpha^{i,j} = \#\{P = (p_1, p_2) \in PD \mid p_1 \leq i \leq j \leq p_2\}$. More details of this method can be seen in [15]. Restricting these functions to a closed interval of $(a, b) \subset \mathbb{R}$ and choosing a uniform

discretization will result in a 2-dimensional feature vector suitable for machine learning classifiers. In this paper, we set $k = 100$ to use the 100 largest such functions in our analysis of mammogram classification.

Persistence image (PI). PI is one of the popular vectorization methods used to transform the topological information contained in PDs into a vector. To construct PI, first rotate PD by $\pi/4$ then turn the rotated PD into a persistent surface via $\Phi : \mathbb{R}^2 \to \mathbb{R}$ and a Gaussian distribution Φ_μ so that $\Phi(\mathrm{PD}) = \sum_{\mu \in PD} w(\mu)\, \Phi_\mu(z)$, where w is a piecewise linear weight function. Finally, the persistent surface Φ is discretized by taking samples over a regular grid.

Persistence binning (P-binning). This approach is one of the simple vectorization methods that relies on counting the number of bars in PBs that intersect with each vertical line $V = 0, 1, 2, \ldots, \omega$. In this paper, we set $\omega = 30$ equidistance vertical lines. Thus, a topological feature vector of size ω was obtained for different dimensions of PBs.

Barcode statistics (P-statistics). The simplest approach to vectorize the space of PBs is to extract statistics directly from PBs. We collected only 10 statistics: average and standard deviation of birth, death and lifespan of bars, median of birth, death and lifespan of bars, and finally the number of bars. The statistics of birth of topological features in the dimension zero of PBs can be ignored as they returned zero by default.

3. Dataset Description and Evaluation Scheme

Two widely used mammogram databases were utilized to test the performance of landmark-based PH for mammogram abnormality classification, which are publicly available. The two datasets are known as Digital Database for Screening Mammogram (DDSM) [17] and Mini Mammographic Image Analysis Society (Mini-MIAS) [18]. Mini-MIAS dataset contains 113 abnormal and 209 normal mammograms of women breasts, which include fatty, granular, calcification, architectural symmetry, and dense cases. DDSM constitutes 2620 mammograms in total, in which 512 mammograms were randomly selected in our experiments with 302 normal cases and 257 abnormal cases. Images in both datasets were cropped region of interest (ROI) images with the size 128-by-128. A number of benchmarking mammographic datasets are available for experimental purposes in which they vary according to certain pre-defined criteria such as type and structure of the digital mammogram, dense, fatty or glandular tissues, noise level in the images, and the number of benign and malignant cases in these datasets. We opted to use Mini-MIAS and DDSM due to the fact that images in both datasets were captured in uncontrolled conditions, so the images contained sufficient noise and low-resolution images. Examples of images from the Mini-MIAS and DDSM datasets can be seen in Figure 5.

Two evaluation metrics that were used are sensitivity (SE), the proportion of breast cancer cases correctly classified as patients having malignant tumors, and specificity (SP), which corresponds to the number of normal breast mammogram cases correctly classified as normal. The accuracy and F1-score is the harmonic mean of precision and recall.

The formula for both sensitivity and specificity is defined as follows:

$$\text{Sensitivity} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}}$$

where true positive (TP) refers to cancer patients truly identified as patients having abnormal breast mammograms, and false negative (FN) refers to breast cancer patients misclassified as negative of having breast cancers.

$$\text{Specificity} = \frac{\text{True negative}}{\text{True negative} + \text{False positive}}$$

where the true negative (TN) refers to the number of truly classified women clear of breast cancer, and false positive (FP) means the number of cases wrongly classified as breast

cancer positive, which in fact are clear of having cancer. The formula for both accuracy and F1-score can be stated as follows:

$$\text{Accuracy} = \frac{\text{TN+ TP}}{\text{TN+ TP+FP+FN}},$$

$$\text{F1} - \text{score} = \frac{2\times\text{TP}}{(2\times\text{TP})+\text{FP+FN}}.$$

The support vector machine (SVM) classifier used to differentiate abnormal mammograms from their normal counterparts optimized all hyperparameters of SVM with a 5-fold cross validation setting.

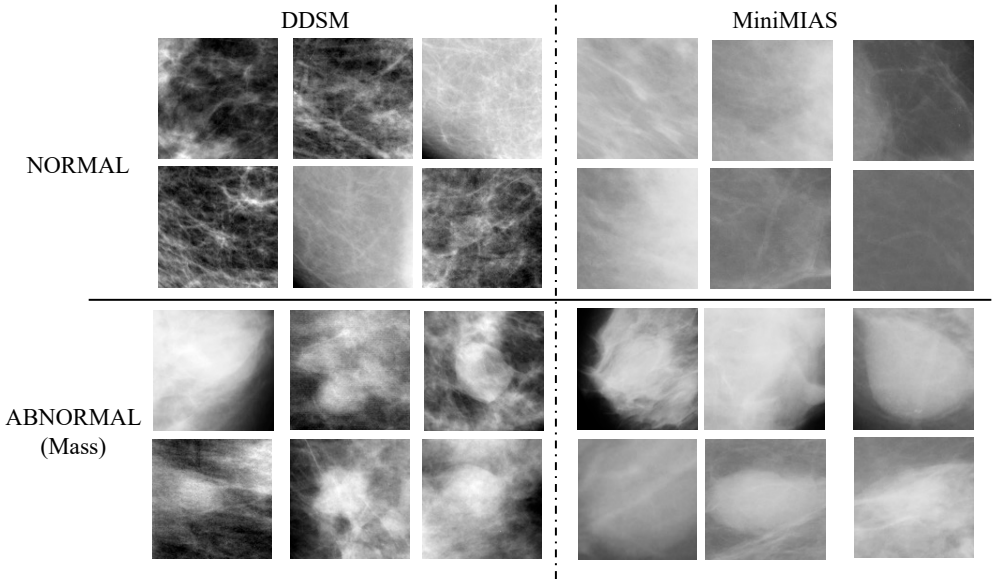

Figure 5. Examples of ROI for normal and abnormal cases from the Mini-MIAS and DDSM datasets.

4. Reproducibility and Implementation Details

In all experiments, we extracted 0-dimensional and 1-dimensional PDs using the Ripser package in python (https://pypi.org/project/ripser/ (accessed on 20 August 2022)). The PI of a resolution 30-by-30, linear weighting function, and the rest of the parameters in default setting were generated using GUDHI library in python (https://pypi.org/project/gudhi/ (accessed on 20 August 2022)). PL was generated with $k = 100$ and the rest of the other parameters with default setting from the GUDHI library. Cubical complex filtration and its corresponding PD was constructed using the GUDHI library in python. SVM classification was performed in MATLAB with standardization and tuning for the optimal hyperparameters. In other words, in each fold, we search for the best kernel among the four kernel options available in MATLAB, which are linear, Gaussian, radial basis function, and polynomial. This means that a linear kernel for the first fold may not be good in the second fold and we may have a case of four different kernels in a 5-fold cross validation. ULBP was implemented from scratch in python to select landmarks. A padding of zero was performed during the process of ULBP landmark selection during 3-by-3 patch scanning of mammograms with an overlap value of 2 between two consecutive patches. The code to reproduce results can be found in the GitHub repository (https://github.com/dashtiali/mammogram-classification (accessed on 1 October 2022)). Full details on how to properly use the code can be found in the GitHub link provided above. A MATLAB version of landmark selection and PH generation and visualization can be found in [19].

5. Experimental Results

In order to classify the mammogram scans, the SVM classifier was trained in a 5-fold cross validation setting based on the topological features. For each image, there was 56 PDs built on 56 point-clouds extracted from the ULBP landmark selection method. There are many approaches in which one can train and test the machine learning classifier for the 56 vectorized PDs generated. We first concatenated the topological features following the seven geometrical groups in ULBP. In other words, features of the eight rotations of each ULBP geometry concatenated for each PD vectorization method. In addition to the seven feature vectors obtained, the PH features extracted in dimension zero and one, which corresponded to the connected components and 1-dimensional cycles. The experimental results obtained from combining topological features in dimension zero and one were better than using either dimension alone.

In Tables 1 and 2, we report on the sensitivity and specificity of the best performing ULBP geometry obtained from the best performing dimension of the PH features and the four PD vectorization methods. Out of the seven ULBP geometries, none of the geometries performance was consistent in both datasets using either of the four vectorization methods utilized. PL with G3 performed better than the other ULBP geometries on DDSM while P-Binning and G7 performed the best on the Mini-MIAS dataset. Combined PH features of dimension zero and one for all 56 ULBP geometries with PL provided 92% sensitivity and 86% specificity for DDSM (see Table 3). The results reported here can be partially compared with that reported in [20], where ULBP and PH were used for mammogram abnormality classification.

The authors in [20] only used binning to vectorize PD with the KNN classifier and they reported the best classification performance of a sensitivity of 86% and specificity of 98% for Mini-MIAS together with an 82% sensitivity and 75% specificity for DDSM. Our results outperformed these results in both datasets.

Finally, in Tables 4 and 5, we report the classification performance of SVM using the cubical complex filtration approach where we used all grayscale pixel values of the mammograms to construct one PD and then the four vectorization methods.

Table 1. The top performing ULBP geometries and PH dimension and all PD vectorizations for DDSM. Avg = average, Std = standard deviation for 5-fold cross-validation using SVM.

Feature Type	Classification Metrics	Avg $\pm$ Std
PD-dim0, 1, P-Binning, and G_5	Sensitivity	85.02 $\pm$ 7.5
	Specificity	77.4 $\pm$ 2.7
	Accuracy	81.57 $\pm$ 4.6
	F1-Score	83.18 $\pm$ 4.8
PD-dim1, P-Statistics, and G_3	Sensitivity	85.1 $\pm$ 4.9
	Specificity	79.7 $\pm$ 6.6
	Accuracy	82.64 $\pm$ 2.7
	F1-Score	84.11 $\pm$ 2.4
PD-dim0, 1, PI, and G_3	Sensitivity	76.4 $\pm$ 9.4
	Specificity	66.9 $\pm$ 7.3
	Accuracy	72.1 $\pm$ 2.8
	F1-Score	74.53 $\pm$ 4.3
PD-dim0, 1, PL, and G_3	Sensitivity	**86.06 $\pm$ 4.8**
	Specificity	**80.9 $\pm$ 4.4**
	Accuracy	**83.7 $\pm$ 4**
	F1-Score	**85.07 $\pm$ 3.7**

Table 2. The top performing ULBP geometries and all PD vectorizations for Mini-MIAS.

Feature Type	Classification Metrics	Avg ± Std
PD-dim0, P-Binning, and G_7	Sensitivity	**97.6 ± 1.5**
	Specificity	**95.5 ± 3.2**
	Accuracy	96.92 ± 1.5
	F1-Score	97.63 ± 1.1
PD-dim0, 1, P-Statistics, and G_5	Accuracy	98.6 ± 2.9
	Specificity	94.6 ± 3.8
	Accuracy	97.27 ± 2.1
	F1-Score	97.9 ± 1.7
PD-dim0, 1, PI, and G_7	Sensitivity	98.1 ± 1.0
	Specificity	94.6 ± 2.1
	Accuracy	96.89 ± 1.1
	F1-Score	97.62 ± 0.9
PD-dim0, 1, PL, and G_7	Sensitivity	97.6 ± 0.1
	Specificity	92.8 ± 6.1
	Accuracy	95.94 ± 2.2
	F1-Score	96.92 ± 1.6

Table 3. Concatenation of all ULBP geometries together with dimension 0 and 1 of the PD for DDSM classification using the top three PD vectorization.

Feature Type	Classification Metrics	Avg ± Std
PD-dim0, 1, and PL	Sensitivity	**92.3 ± 4**
	Specificity	**86.5 ± 3**
	Accuracy	89.62 ± 1.4
	F1-Score	90.56 ± 1.4
PD-dim0, 1, and P-Binning	Sensitivity	85.1 ± 6
	Specificity	82.6 ± 4
	Accuracy	83.57 ± 4.3
	F1-Score	84.73 ± 4.2
PD-dim0, 1, and P-Statistics	Sensitivity	87.6 ± 4
	Specificity	82.3 ± 4
	Accuracy	84.62 ± 1.5
	F1-Score	85.93 ± 1.6

Table 4. Cubical complex performance results for the Mini-MIAS dataset using four different vectorization methods and three homology dimensions.

Feature Type	Sensitivity (Avg ± STD)	Specificity (Avg ± STD)	Accuracy (Avg ± STD)	F1-Score (Avg ± STD)
P-Binning and PD-dim0	99.02 ± 2.18	2.51 ± 3.65	65.17 ± 1.3	78.68 ± 0.9
P-Binning and PD-dim1	99.02 ± 1.34	0.91 ± 2.03	64.6 ± 0.7	78.41 ± 0.5
P-Binning and PD-dim0, 1	98.54 ± 2.18	2.51 ± 3.65	64.86 ± 1.5	78.45 ± 1
P-Statistics and PD-dim0	98.54 ± 1.34	92.15 ± 3.25	96.29 ± 0.8	97.18 ± 0.6
P-Statistics and PD-dim1	**98.09 ± 1.07**	**96.47 ± 1.99**	**97.52 ± 0.8**	**98.09 ± 0.6**
P-Statistics and PD-dim0, 1	**98.58 ± 1.3**	**94.76 ± 1.54**	**97.24 ± 1.2**	**97.88 ± 0.9**
PI and PD-dim0	88.68 ± 5.53	78.04 ± 13.48	84.95 ± 6.2	88.46 ± 4.7
PI and PD-dim1	90.95 ± 4.55	84.98 ± 3.81	88.86 ± 3.6	91.34 ± 2.9
PI and PD-dim0,1	93.83 ± 2.53	91.24 ± 2.67	92.92 ± 2	94.49 ± 1.6
PL and PD-dim0	93.39 ± 3.23	69.96 ± 12.29	85.17 ± 5.8	89.16 ± 4.2
PL and PD-dim1	89.7 ± 8.78	78.51 ± 8.56	85.81 ± 5.3	88.99 ± 4.6
PL and PD-dim0, 1	95.43 ± 6.23	80.22 ± 11.68	90.13 ± 3.8	92.59 ± 2.9

Table 5. Cubical complex performance results for the DDSM dataset using four different vectorization methods and three homology dimensions.

Feature Type	Sensitivity (Avg ± STD)	Specificity (Avg ± STD)	Accuracy (Avg ± STD)	F1-Score (Avg ± STD)
P-Binning and PD-dim0	57.08 ± 15.13	62.94 ± 10	59.78 ± 5.5	59.69 ± 9.7
P-Binning and PD-dim1	68.6 ± 17.5	62.26 ± 12.89	65.69 ± 4.2	67.44 ± 8.3
P-Binning and PD-dim0, 1	69.16 ± 19.52	51.14 ± 22.9	60.87 ± 7.5	64.72 ± 8.8
P-Statistics and PD-dim0	82.08 ± 7.57	80.57 ± 5.58	81.38 ± 4.4	82.55 ± 4.4
P-Statistics and PD-dim1	80.13 ± 8.5	87.19 ± 3.88	83.37 ± 5.5	83.73 ± 5.9
P-Statistics and PD-dim0, 1	**86.03 ± 8**	**81.72 ± 4.89**	**84.05 ± 4.2**	**85.23 ± 4.3**
PI and PD-dim0	61.01 ± 9	81.69 ± 7.85	70.52 ± 6	68.89 ± 7.2
PI and PD-dim1	71.62 ± 10.01	78.19 ± 5.35	74.65 ± 6.2	75.07 ± 7.2
PI and PD-dim0, 1	73.19 ± 6.18	73.95 ± 6.48	73.54 ± 4.2	74.87 ± 4.2
PL and PD-dim0	74.46 ± 7.17	71.54 ± 7.86	73.12 ± 5.3	74.89 ± 5.1
PL and PD-dim1	81.13 ± 6.06	73.53 ± 7.34	77.63 ± 5.5	79.66 ± 5
PL and PD-dim0, 1	83.74 ± 4.72	77.01 ± 5.03	80.65 ± 4.7	82.37 ± 4.3

It can be seen that by using cubical complexes, we can obtain 98% and 94% for sensitivity and specificity, respectively, for the Mini-MIAS dataset and up to 86% of sensitivity and 81% specificity for DDSM. P-Statistics performed better than the other three vectorization methods using cubical complexes, which was not the case using landmark-based VR filtration. Using our proposed landmark based approach, one geometry alone (G7) achieved roughly the same performance for Mini-MIAS where we only used a small portion of the mammogram scan pixel values. For DDSM, we outperformed the cubical complexes if we concatenated the topological features from all geometries and obtained almost the same performance using one geometry (i.e., G3).

6. Discussion and Future Work

This study introduced a distributed method of constructing 56 PDs based on automatically extracted landmarks from breast mammograms. In general, we found that a small set of pixel landmarks was enough to detect abnormality in breast mammograms such as G3 in DDSM and G7 in the Mini-MIAS dataset. Computing the 56 PDs can be conducted in a distributed manner, which is crucial for large scale datasets. Instead of building one PD using cubical complexes, our approach provides a localized PD representation that conveys topological features linked to different types of mammogram texture distribution. Different PD vectorizations were examined where we found that it was good practice to try more than one method, as a single approach may not consistently perform well on different datasets. This work is the first step toward a more comprehensive study for different approaches of PD vectorization in medical imaging because we concluded that different types of vectorization methods affect the performance greatly, as can be seen from Table 1 to Table 5. Until now, to the best of our knowledge, there is no comprehensive analysis or a roadmap to select suitable vectorization method(s) for medical image analysis or any other image modalities. On the other hand, it is not an easy task to search the rich literature of PH vectorizations and pick the correct method suitable for the problem at hand. Nonetheless, based on the findings in this work, at this stage, we are not advising the use of a single vectorization method, as this could lead to misleading performances.

Furthermore, our analysis showed that the proposed landmark based PH could outperform classical approaches of building topology from digital images such as cubical complexes. This points to the fact that a small set of pixel value landmarks that correspond to different type of textures can be used to differentiate malignant mammogram scans from benign scans. This is particularly useful when the medical image dimensions (number of rows and columns) are very high, and using the entire pixel values is time consuming, or downsampling may result in the loss of critical medical information.

The only limitation of this work is the increase in the dimensions of feature vectors when combining more than one ULBP geometry, as was the case when we concatenated all

ULBP geometries to boost the classification performance for the DDSM dataset in Table 3. Aggregating PDs before vectorization is one approach to address this limitation in future. Future work will also focus on using other texture methods as landmark selection procedures such as center-symmetric LBP or small image patches of high density. Furthermore, testing the proposed ULBP based PH on other medical image modalities such as ultrasounds and other types of disease is included in our list of future works.

Author Contributions: Conceptualization, A.A. and R.R.; Data curation, A.A., D.A. and R.R.; Formal analysis, T.M.; Methodology, A.A., T.M. and R.R.; Software, A.A. and D.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No data is available.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Carlsson, G. Topology and Data. *Bull. Am. Math. Soc.* **2009**, *46*, 255–308. [CrossRef]
2. Cohen-Steiner, D.; Edelsbrunner, H.; Harer, J. Stability of Persistence Diagrams. *Discrete Comput. Geom.* **2007**, *37*, 103–120. [CrossRef]
3. WHO. *International Agency for Research on Cancer*; WHO: Geneva, Switzerland, 2020.
4. Ojala, T.; Pietikäinen, M.; Harwood, D. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognit.* **1996**, *29*, 51–59. [CrossRef]
5. Xu, Q.; Yang, J.; Ding, S. Texture Segmentation using LBP embedded Region Competition. *ELCVIA Electron. Lett. Comput. Vis. Image Anal.* **2005**, *5*, 41–47. [CrossRef]
6. Heikkilä, M.; Pietikäinen, M.; Schmid, C. Description of interest regions with local binary patterns. *Pattern Recognit.* **2009**, *42*, 425–436. [CrossRef]
7. Abbasi, S.; Tajeripour, F. Detection of brain tumor in 3D MRI images using local binary patterns and histogram orientation gradient. *Neurocomputing* **2017**, *219*, 526–535. [CrossRef]
8. Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. [CrossRef]
9. Ahonen, T.; Hadid, A.; Pietikäinen, M. Face recognition with local binary patterns. *Lect. Notes Comput. Sci.* **2004**, *3021*, 469–481.
10. Nanda, V.; Sazdanović, R. Simplicial models and topological inference in biological systems. In *Natural Computing Series*; Springer: Berlin/Heidelberg, Germany, 2014; Volume 48, pp. 109–141.
11. Otter, N.; Porter, M.A.; Tillmann, U.; Grindrod, P.; Harrington, H.A. A Roadmap for the Computation of Persistent Homology. *EPJ Data Sci.* **2017**, *6*, 17. [CrossRef] [PubMed]
12. Garin, A.; Tauzin, G.A. Topological "Reading" Lesson: Classification of MNIST using TDA. In Proceedings of the 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), Boca Raton, FL, USA, 16–19 December 2019; pp. 1551–1556.
13. Turner, K.; Mileyko, Y.; Mukherjee, S.; Harer, J. Fréchet Means for Distributions of Persistence Diagrams. *Discret. Comput. Geom.* **2014**, *52*, 44–70. [CrossRef]
14. Adams, H.; Emerson, T.; Kirby, M.; Neville, R.; Peterson, C.; Shipman, P.; Chepushtanova, S.; Hanson, E.; Motta, F.; Ziegelmeier, L.; et al. Persistence Images: A Stable Vector Representation of Persistent Homology. *J. Mach. Learn. Res.* **2017**, *18*, 218–252.
15. Bubenik, P. Statistical Topological Data Analysis using Persistence Landscapes. *J. Mach. Learn. Res.* **2015**, *16*, 77–102.
16. Asaad, A.T.; Rashid, R.D.; Jassim, S.A. Topological image texture analysis for quality assessment. In Proceedings of the SPIE—The International Society for Optical Engineering, Anaheim, CA, USA, 9–13 April 2017; Volume 10221.
17. Heath, M.; Bowyer, K.; Kopans, D.; Moore, R.; Kegelmeyer, P. The digital database for screening mammography. In Proceedings of the Fifth International Workshop on Digital Mammography, Toronto, ON, Canada, 11–14 June 2001; pp. 212–218.
18. Suckling, J.; Parker, J.; Dance, D.; Astley, S.; Hutt, I.; Boggis, C.; Ricketts, I.; Stamatakis, E.; Cerneaz, N.; Kok, S.; et al. *Mammographic Image Analysis Society (MIAS) Database*; The University of Cambridge: Cambridge, UK, 2015. Available online: https://www.repository.cam.ac.uk/handle/1810/250394 (accessed on 20 September 2012).
19. Ali, D.; Asaad, A. DAAR Topology: A Software to Build Topological Features from Images. Available online: https://github.com/daartopology/DAAR-Topology (accessed on 16 October 2021).
20. Asaad, A. *Persistent Homology Tools for Image Analysis*; The University of Buckingham: London, UK, 2020.