September 2024
Volume 5 Nomor 2

p-ISSN 2723-0325 e-ISSN 2723-0333

TENSOR

Pure and Applied Mathematics Journal

PROGRAM STUDI MATEMATIKA
JURUSAN MATEMATIKA
FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS PATTIMURA

TENSOR
Pure and Applied Mathematics Journal

is an international academic open-access journal that gains a foothold in mathematics, and its applications are issued twice a year. The focus is to publish original research and review articles on all aspects of pure and applied Mathematics. Editorial board members of the Journal and reviewers will review submitted papers. All submitted articles should report original, previously unpublished research results, experimental or theoretical, and will be peer-reviewed. Articles submitted to the journal should meet these criteria and must not be under consideration for publication elsewhere. Manuscripts should follow the journal template and are subject to both review and editing.

Published by:

Department of Mathematics,
Faculty of Mathematics and Natural Sciences,
Pattimura University.
Ambon
2024

TENSOR
Pure and Applied Mathematics Journal
Volume 5 Number 2 | September 2024

Publisher
Department of Mathematics,
Faculty of Mathematics and Natural Sciences,
Pattimura University, Ambon, Indonesia

Editorial Address
Program Studi Matematika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Unversitas Pattimura
Jln. Ir. M. Putuhena, Kampus Unpatti, Poka - Ambon 97233, Provinsi Maluku, Indonesia
Contact : +62 82397854220
Email : tensormathematics@gmail.com

Volume 5 Number 2 | September 2024

TENSOR
Pure and Applied Mathematics Journal

TENSOR Pure and Applied Mathematics Journal

September 2024 | Vol. 5 | No. 2 pp. 79–92

p-ISSN: 2723-0325 e-ISSN: 2723-0333

Exploring the Lazy Witness Complex for Efficient Persistent Homology in Large-Scale Data

Mst Zinia Afroz Liza[1,*], Md. Al-Imran[1], Md. Morshed Bin Shiraj[1], Tozam Hossain[2], Md. Masum Murshed[1] and Nasima Akhter[1]

[1] Department of Mathematics, University of Rajshahi, Rajshahi-6205, Bangladesh.
[2] Department of Mathematics, Bangamata Sheikh Fojilatunnesa Mujib Science and Technology University, Bangladesh.
[*]Email: mbshiraj@gmail.com

Manuscript submitted : October 2024;
Accepted for publication : November 2024.
doi: https://doi.org/10.30598/tensorvol5iss2pp79-92

Abstract: In this paper, topological data analysis (TDA) techniques have been explored, with a focus on the selection of the Witness Complex and Persistent Homology of some nested families of Lazy Witness Complex as approximations for analyzing complex datasets. The Witness Complex was chosen for its efficiency and scalability, as it constructs a simplicial complex using landmark points, reducing computational load compared to methods like the Vietoris-Rips and Čech complexes. This makes it suitable for large, high-dimensional datasets, accurately representing the dataset's intrinsic geometry even with varying data densities. Persistent Homology was then reviewed with the aim of calculating it on some nested families of the Witness Complex. Subsequently, the nested families of the Lazy Witness Complex were introduced mathematically, with an example of the entire construction process for a well-known point cloud dataset. For this purpose, 50 points were generated randomly from a circle, and persistent diagrams of the point cloud data were analyzed to understand and compare the behavior among the approximations of the Witness Complex after choosing 10 landmarks using the Maxmin method. Since the families are nested, the filtration process became faster for each successive family, thus reducing computational complexity. For all three cases $v = 0,1,2$, the persistent barcodes indicated the same shape as the dataset. This study may help in choosing the suitable family of the Witness Complex over Persistent Homology to balance computational feasibility with topological accuracy, enabling efficient handling of large datasets while preserving important topological features. This approach allows for extracting meaningful insights from complex data while effectively managing computational resources.

2020 Mathematical Subject Classification: 55-08, 68-04
Keywords: Landmark Points, Persistent Homology, Simplicial Complex, Witness Complex.

1. Introduction

In the ever-changing world of data analysis, the search for methods that can accurately grasp the intricate structures and patterns hidden within complex datasets is crucial. Topological data analysis (TDA) is one such method that stands out for its ability to reveal the underlying geometric and topological features of data, going beyond traditional statistical approaches. Central to TDA is the idea of simplicial complexes, which represent how data points are connected in a geometrically meaningful way. However, simplicial complexes may not work well with datasets that are high-dimensional or noisy.

In response to these challenges, witness complexes emerge as a promising alternative, offering a robust and efficient approach to capturing the topological essence of complex data. Witness complexes, a concept pioneered by Vin de Silva and Gunnar Carlsson [9], provide a means to construct simplicial complexes based not only on the original data points but also on additional "witness points" strategically placed within the data domain. This augmentation serves to enhance the fidelity of the resulting complex, enabling a more comprehensive representation of the underlying topology.

Witness complexes are defined for data sets in any metric space, not necessarily in Euclidean space. Witness complex helps in simplifying the data while still preserving its essential topological features. By considering the interaction between data points and landmark points, it provides a more computationally tractable representation of the data for persistent homology calculations. This simplification is useful of understanding the persistent homology of large and complex data sets.

In [9] researchers construct a simplicial complex named Witness complex which is smaller and provides better pictures of the homology with less noise than the other complexes. Researchers gives a reconstruction algorithm that the restricted Delaunay triangulation can be replaced by the Witness complex which is applicable in any metric space in [14]. The structural properties of the geodesic Witness complex and analog of the usual Witness complex in the intrinsic metric are discussed in [12]. The Witness simplicial Variational auto-encodes (VAEs) as an extension of the simplicial auto-encoder to the variational setup using a Witness complex for computing the simplicial regularization is proposed in [19]. Witness complex is also used in [5] for dynamical analysis of time series. In the meantime, Persistent homology is used in complex network science [4], nuclear collision [15], and analysis of brain transcriptome data in autism [21]. It is also used in natural language processing on linguistic data such as text corpora or speech signals [22].

The computation of Persistent homology is an open area with numerous important and fascinating challenges. The field of Persistent homology captures the emergence and disappearance of topological traits throughout a filtration constructed from a dataset [13], succinctly summarizing data's topological features through persistent diagrams or barcodes. This summarization is pivotal for tracking changes and facilitating the analysis of data across multiple scales, although it presents challenges due to the complex nature of the data structure associated with topological features, complicating learning tasks. These persistent diagrams are subsequently transformed into metric spaces with additional structure conducive to machine learning endeavors [2]. With the remarkable success of deep learning in computer vision tasks [18], deep networks capable of handling barcodes have emerged [16]. In [6], persistent diagrams were integrated with neural network classifiers for graph classification tasks. Moreover, persistent barcodes were employed for classifying brain activation patterns in rs-fMRI video frames [10,11]. The topological and geometric structures underlying data are often represented as point clouds, with recent discussions focusing on multiclass classification of point cloud datasets [17].

Simplicial complexes such as Čech, Rips, or α-shapes are standard tools for approximating the topological features of an underlying space. These complexes typically use a vertex set that is as large as the original point cloud data, leading to constructions that are sometimes manageable but often highly inefficient. This inefficiency arises because the homotopy types of the underlying structures can usually be represented with significantly fewer vertices, as noted by [9]. The challenge becomes especially pronounced with Rips and Čech complexes, which are notorious for their rapid growth in size. For instance, a Rips complex formed from n points in $\mathbb{R}^d$ can generate $\Omega(n^d)$ simplices [7], leading to an overwhelming number of simplices as the data size increases. To put this in perspective, when constructing a Rips or Čech complex from even a few thousand points, the resulting structure may contain millions of triangles, rendering it impractical for computation or analysis. This issue is particularly acute in lower dimensions, such as three, where the complexity of the resulting simplicial complexes quickly escalates beyond what can be handled efficiently.

To address these challenges, the witness complex presents a more scalable alternative. Instead of using the entire dataset, the witness complex employs a subsampling technique by selecting a strategically chosen subset of "landmark" points. These landmark points serve as a representative sample of the original data, and the complex is then built around them, significantly reducing the number of vertices and simplices involved (see Fig. 1). This approach is inspired by the Delaunay complex in Euclidean space, where the focus is on capturing the essential topological features with a much smaller, more manageable structure. By doing so, the witness complex not only retains the critical topological information but also makes the computational process more efficient and feasible for large-scale data sets. This method represents a significant advancement in the practical application of simplicial complexes for topological data analysis, enabling the study of larger and more complex data sets that would otherwise be intractable.

Since Witness complex provides more robust calculation for homology than any other methods like Vietoris-Rips complex and Čech complex (see example 3.1 in [9]), so by using Witness complex computing persistent homology of big data will be easy. In this study, the computation of families of Lazy Witness Complexes has been reviewed, and the efficiency of the Lazy Witness Complex has been demonstrated by using 50 random data points from a circle as a representative of large-scale data.

2. Mathematical Background

2.1. Simplicial complex

Definition: A set of simplices constitutes what is known as a simplicial complex $\mathcal{K}$ such that,

(a) If $\mathcal{K}$ contains a simplex σ, then it contains every face of σ.

(b) If two simplices in $\mathcal{K}$ intersects, then their intersection is either empty or a face of each of them [20].

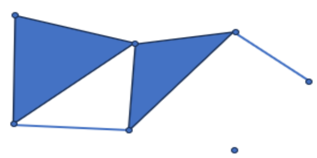

Fig. 1. An example of a Simplicial Complex.

Simplicial complexes (Fig. 1) are fundamental structures in topological data analysis, used to represent the shape of data in a combinatorial way.

2.2. Vietoris-Rips Complex

Definition: Consider a metric space X and a subset $S \subseteq X$, where the vertex set is represented by S. Let's select a scale parameter $r \geq 0$. Then, the Rips complex, denoted as $\text{Rips}(S, r)$ is an abstract simplicial complex defined as follows:

$$\text{Rips}(S, r) = \{\text{finite } \sigma \subseteq S | \text{Diam}(\sigma) \leq r\}$$

Here, $\text{Diam}(\sigma)$ denotes the diameter of the set σ.

Vietoris-Rips complexes, also known as Rips complexes. In this context, the condition $\text{Diam}(\sigma) \leq r$ implies that the distance from one vertex to another within σ does not exceed r. The Čech complex, applicable to any finite metric space, finds an approximation through the Vietoris-Rips complex. While computing it for extensive point datasets can be resource-intensive, it's generally more manageable compared to the Čech complex which can be cleared in Fig. 2(a). Here, three balls intersect pairwise without a common intersection, forming a 2-simplex in the Rips complex $[abc]$.

2.3. Čech Complex

Definition: Let $X = \{v_1, v_2, \dots, v_n\} \subseteq \mathbb{R}^n$ and $r_1 > 0$. The Čech complex, denoted as Čech(X, r_1), is a simplicial complex abstractly constructed with the vertex set $\{v_1, v_2, \dots, v_n\}$. In this complex, a subset $\{v_{i_1}, v_{i_2}, \dots, v_{i_k}\}$ belongs to Čech(X, r_1), if and only if the intersection $B(v_{i_1}, r_1) \cap B(v_{i_2}, r_1) \cap \dots \cap B(v_{i_k}, r_1) \neq \emptyset$.

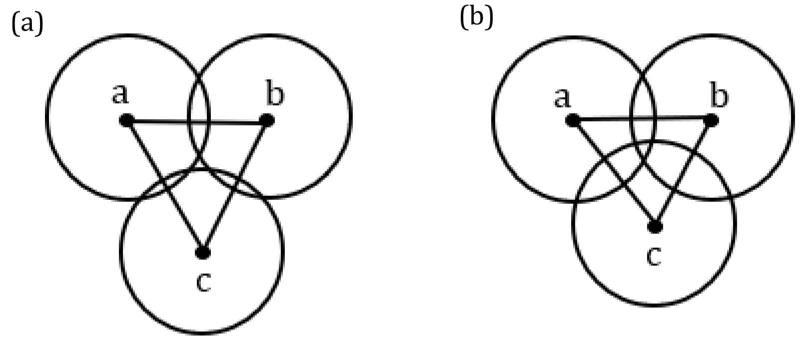

Fig. 2. Examples of (a) Vietoris-Rips and (b) Čech complex.

The Čech complex constructs simplices by considering the intersection of balls centered at points within a given radius in the metric space. Whenever this intersection contains at least one point, a simplex is formed. In simpler terms, a simplex in the Čech complex represents a group of points that are all within a set distance of each other. In the Fig. 2(b) provided, there are three balls intersecting pairwise, sharing a common intersection. This configuration forms a 2-simplex $[abc]$ within the Čech complex.

2.4. Filtration of Simplicial Complex

Definition [23]: Consider K as a simplicial complex. A discrete filtration (Fig. 3) of K refers to a sequence of successive subcomplexes

$$K_1 \leq K_2 \leq \cdots \leq K_m = K.$$

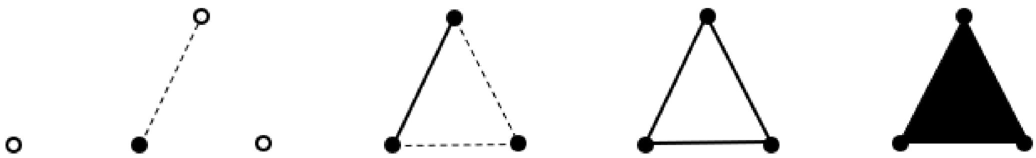

Fig. 3. Filtration of a complex.

2.5. Persistent Homology

Definition [23]: Suppose K forms a simplicial complex, $\mathbb{F}$ denotes a field, and $q \in \mathbb{Z}^+$. Given a filtration

$$K_1 \leq K_2 \leq \cdots \leq K_m = K$$

of K, the associated q-dimensional persistent homology groups with coefficients in $\mathbb{F}$ represent the images of the maps

$$(i_{s,t})_*: H_q(K_s; \mathbb{F}) \to H_q(K_t; \mathbb{F}), \text{ for all } 0 \leq s \leq t \leq m.$$

For a filtered complex, each i-th complex K^i is associated with boundary operators δ_k^i, matrices M_k^i, and groups C_k^i, Z_k^i, B_k^i, and H_k^i $\forall i$ & $k \geq 0$. The p-persistent kth homology group of K^i is given by:

$$H_k^{i,p} = \frac{Z_k^i}{B_k^{i+p} \cap Z_k^i}.$$

3. Witness Complex

3.1. Definition of W(D)

According to [9], let D be an $n \times N$ matrix where each entry contains a non-negative value representing the distance between n landmark points and N data points in a given dataset. The (strict) witness complex, denoted as $W_\infty(D)$, is constructed with a vertex set $\{1,2,\dots,n\}$, following these specific rules:

Edges: An edge $\sigma = [ab]$ is included in $W_\infty(D)$ if and only if there exists a data point, indexed by i (where $1 \leq i \leq N$), such that the distances $D(a,i)$ and $D(b,i)$ are the two smallest values in the i-th column of D, regardless of their order.

Simplices: By extending this concept through induction on p, suppose all the faces of a p-simplex $\sigma = [a_0 a_1 \cdots a_p]$ are already present in $W_\infty(D)$. The simplex σ itself is included in $W_\infty(D)$ if and only if there exists a data point indexed by i (where $1 \leq i \leq N$) such that the entries $D(a_0,i), D(a_1,i), \dots, D(a_p,i)$ correspond to the smallest $p+1$ values in the i-th column of D, ordered in some manner.

In both of these scenarios, the index i acts as a witness, confirming the existence of the simplex σ within the complex.

Example: Consider the 3×6 matrix:

$$\begin{bmatrix} D(v_1,v_1) & D(v_1,v_2) & D(v_1,v_3) & D(v_1,v_4) & D(v_1,v_5) & D(v_1,v_6) \\ D(v_2,v_1) & D(v_2,v_2) & D(v_2,v_3) & D(v_2,v_4) & D(v_2,v_5) & D(v_2,v_6) \\ D(v_3,v_1) & D(v_3,v_2) & D(v_3,v_3) & D(v_3,v_4) & D(v_3,v_5) & D(v_3,v_6) \end{bmatrix}$$

If we consider v_1 and v_2, they are the smallest entries compared to v_3 with respect to the witness v_5. Therefore, the simplex $[v_1 v_2]$ will be formed.

Again, consider the 4×6 matrix:

$$\begin{bmatrix} D(v_1,v_1) & D(v_1,v_2) & D(v_1,v_3) & D(v_1,v_4) & D(v_1,v_5) & D(v_1,v_6) \\ D(v_2,v_1) & D(v_2,v_2) & D(v_2,v_3) & D(v_2,v_4) & D(v_2,v_5) & D(v_2,v_6) \\ D(v_3,v_1) & D(v_3,v_2) & D(v_3,v_3) & D(v_3,v_4) & D(v_3,v_5) & D(v_3,v_6) \\ D(v_4,v_1) & D(v_4,v_2) & D(v_4,v_3) & D(v_4,v_4) & D(v_4,v_5) & D(v_4,v_6) \end{bmatrix}$$

If we consider v_1, v_2 and v_3, they are the smallest entries compared to v_6 with respect to the witness v_5. Therefore, the simplex $[v_1 v_2 v_3]$ will be formed.

3.2. Weak and Strong Witness

Definition [8]: Let $L \in \mathbb{R}^d$ be a finite collection of points, and $l_0, l_1, \cdots, l_p \in L$. Then a simplex $\sigma = [l_0 l_1 \cdots l_p]$ is weakly-witnessed by $x \in \mathbb{R}^d$ if $d(x,p) \leq d(x,q)$ for every $p \in \{l_0, l_1, \cdots, l_p\}$ and $q \in L \setminus \{l_0, l_1, \cdots, l_p\}$. Equivalently, $x \in \mathbb{R}^d$ serves as a weak-witness for σ concerning L if and only if $d(x,p) \leq d(x,q)$. In simpler terms, the $p+1$ nearest neighbors of x in L are $l_0, l_1, \cdots, l_p$. x is considered as a strong-witness if and only if $d(x,l_0) = \cdots = d(x,l_p)$.

3.3. Lazy Witness Complex

Definition: Similar to the Rips complex, there exists a "lazy" version of the witness complex. We formally define $W_1(D)$ containing $W_\infty(D)$ as follows: $W_1(D)$ shares the same 1-skeleton as $W_\infty(D)$. A p-simplex $\sigma = [a_0 a_1 \cdots a_p]$ is part of $W_1(D)$ if and only if all its edges are in $W_1(D)$.

In simpler terms, $W_1(D)$ is the largest simplicial complex that has the same vertices and edges as $W_\infty(D)$. In practical applications, $W_\infty(D)$ is rarely used due to its more complicated computation, and we denote

$W(D)$ to mean $W_1(D)$.

3.3.1. Filtration of Lazy Witness Complex

The filtration of a Lazy Witness Complex involves incrementally adding simplices based on the "lazy" construction principle, where new simplices are only added if they contribute to the connectivity of the complex. It utilizes a set of landmarks to define neighborhoods and builds simplices based on the pairwise distances between these landmarks and data points (see Fig. 4). This incremental approach allows for efficient construction of a simplicial complex that captures the topological features of the underlying data.

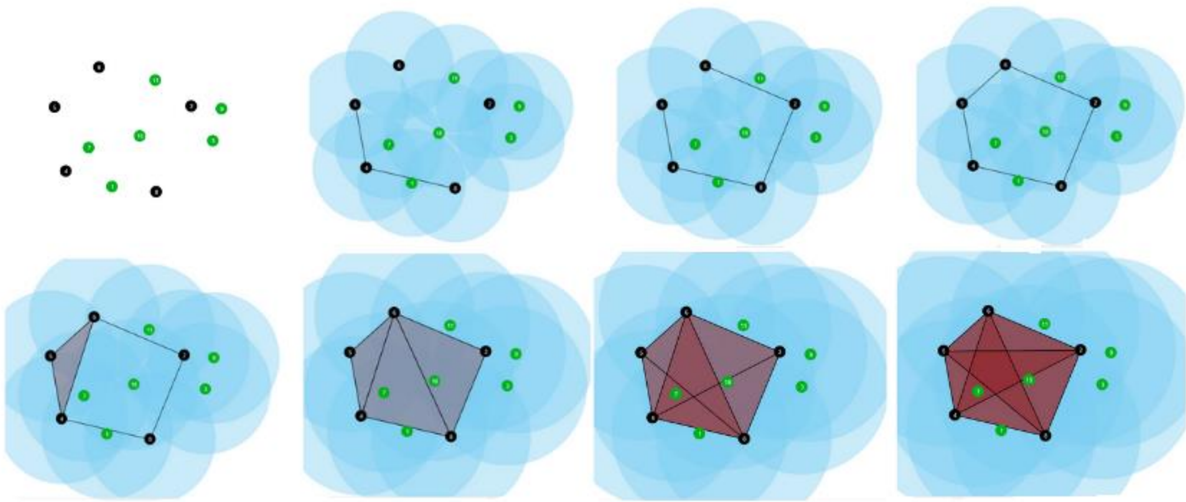

Fig. 4. Filtration of Lazy Witness complex using [3].

3.4. Landmark Selection [9]

We suggest acquiring the landmark set through the Maxmin method:

Define Z as the set of data points. Initially, select l_1 randomly from Z. Subsequently, choose l_k such that $max\{min\{D(l_1, v), D(l_2, v), \cdots, D(l_{k-1}, v)\}\}$, for all $v \in Z \setminus \{l_1, l_2, \ldots\ldots, l_{k-1}\}$.

Maxmin generally results in more evenly spaced landmarks, though it often selects points at the extremes, as demonstrated in Fig. 5.

The number of landmarks should be determined by establishing a minimum ratio of N/n. While we don't have a precise rule for this lower limit, a ratio of $N/n \geq 20$ tends to work well for data sampled from a two-dimensional surface.

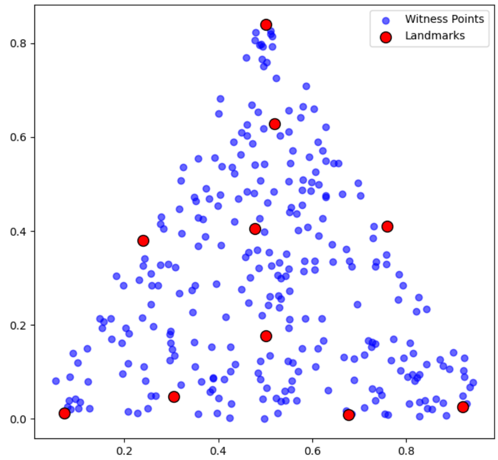

Fig. 5. Selection of Landmark points using Maxmin method.

4. Methods

According to [9], consider a matrix D of dimensions $n \times N$, where each element $D(a,i)$ represents a non-negative distance between n landmark points $\{1,2,\dots,n\}$ and N data points. The matrix serves as the foundation for constructing a series of nested simplicial complexes, denoted by $W(D;R,v)$, where R is a filtration parameter that varies over the interval $[0,\infty]$. The parameter v is a non-negative integer that controls the selection of simplices within the complex. The cases where $v = 0,1,2$ are particularly significant because they represent different levels of strictness in the selection process, impacting the complexity and granularity of the resulting simplicial complexes.

The construction of the complex $W(D;R,v)$ follows these rules:

Vertex Set: The vertex set of the simplicial complex $W(D;R,v)$ is defined as $\{1,2,\dots,n\}$, representing the landmark points.

Case $v = 0$*:* In this scenario, the value m_i for each data point i (ranging from 1 to N) is set to 0. This implies that no additional flexibility is allowed when determining whether simplices should be included in the complex, leading to a more rigid structure where only the most immediate connections (those with the smallest distances) are considered.

Case $v > 0$*:* Here, m_i is defined as the v-th smallest value in the i-th column of D. This allows for a broader inclusion of simplices, as it introduces a level of tolerance by considering not just the smallest distances but the v-th smallest ones, enabling the complex to capture more intricate relationships between data points.

Edges: An edge $\sigma = [ab]$ is included in the complex $W(D;R,v)$ if there exists a data point i (acting as a witness) such that the maximum distance between the landmark points a and b and this data point i does not exceed the sum of the filtration parameter R and the tolerance value m_i. Mathematically, this condition is expressed as $\mathrm{Max}(D(a,i), D(b,i)) \leq R + m_i$.

Higher-Dimensional Simplices: A p −simplex $\sigma = [a_0 a_1 \dots a_p]$ is included in the complex if all of its edges satisfy the condition outlined for edges. This implies that there must be a witness i such that the maximum distance between any vertex of the simplex and the witness is within the allowed range, $\mathrm{Max}\left(D(a_0,i), D(a_1,i), \dots, D(a_p,i)\right) \leq R + m_i$.

This inductive condition ensures that higher-dimensional features in the data are only included in the complex if their constituent lower-dimensional features are also present.

Computation of Persistent Homology

To analyze the evolution of the topological features of the data as the filtration parameter R changes, persistent homology is computed over the interval $R \in [0, r]$. The goal is to identify and track the birth and death of topological features, such as connected components, loops, and voids, across different levels of the filtration. This process is algorithmically implemented as follows:

i. Matrix Construction: Start by computing an $n \times n$ matrix E, where each off-diagonal element $E(i,j)$ records the time $R_{[ij]}$ at which the edge $[ij]$ first appears in the complex. This matrix serves as a crucial tool for determining the inclusion of higher-dimensional simplices.

ii. Simplex List Generation: Generate a list of simplices that appear at or before a specified filtration value r. This list includes all vertices, edges, and higher-dimensional simplices that meet the inclusion criteria for the given R.

iii. Appearance Time Calculation: For each simplex σ, determine its appearance time R_σ as the maximum of the appearance times of its edges. This ensures that the complex respects the nested structure, where a simplex can only be included if all of its lower-dimensional faces are already present.

iv. Tracking Topological Features: By tracking the birth and death of these simplices as R increases, one can construct the persistent homology of the data. This captures the robustness of topological features across scales, distinguishing between features that persist across a wide range of R (indicating significant structure) and those that appear only briefly (often corresponding to noise).

This method of constructing and analyzing the simplicial complex $W(D; R, v)$ provides a powerful framework for understanding the underlying topological structure of the data, with applications ranging from data analysis to machine learning and beyond. The flexibility of the parameter v allows researchers to tailor the analysis to different levels of granularity, making this approach widely applicable to various types of data. This can be cleared in Fig. 6.

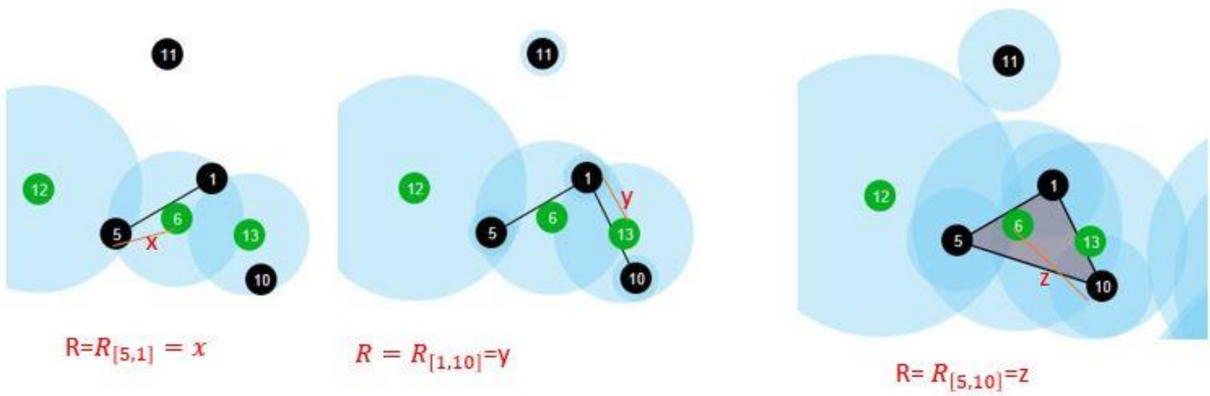

Fig. 6. Time of appearance.

Here $R_{[5,1,10]} = max\{R_\tau : \tau \text{ is an edge of } [5,1,10]\}$
$= R_{[5,10]}$
$= z$

5. Results

We first randomly selected 50 points from a 2D circle in the point cloud and then identified 10 landmark points using the Maxmin method (see Fig. 7). Next, we constructed the Lazy Witness Complex and calculated its persistence to visualize its barcode, thereby identifying its topological features.

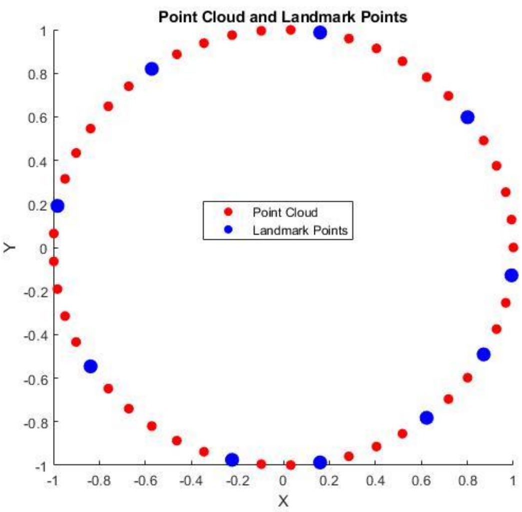

Fig. 7. Persistence barcodes in the case of $v = 0$.

We constructed Lazy Witness Complexes using [1]. Here are the 50 values forming the lazy witness complexes for $v = 0,1,2$ as shown in the Figs. 8-10.

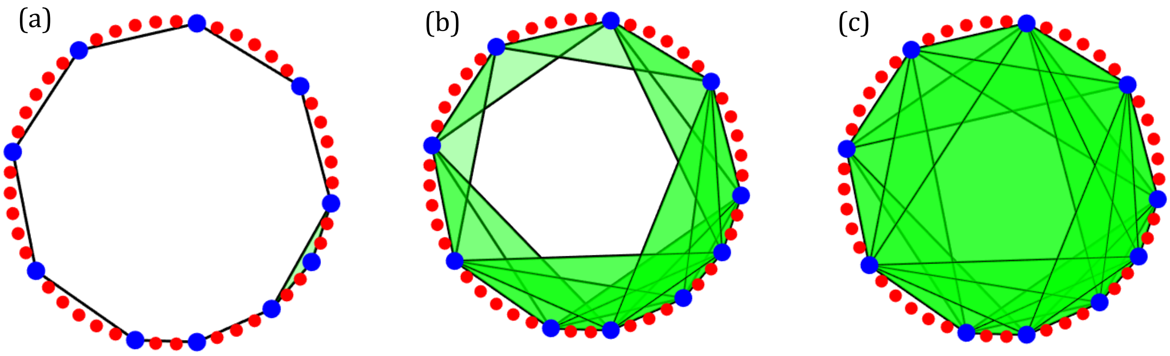

Fig. 8. Lazy Witness Complex for ν = 0: (a) when R = 0.36, (b) when R = 1.03, and (c) when R = 1.08.

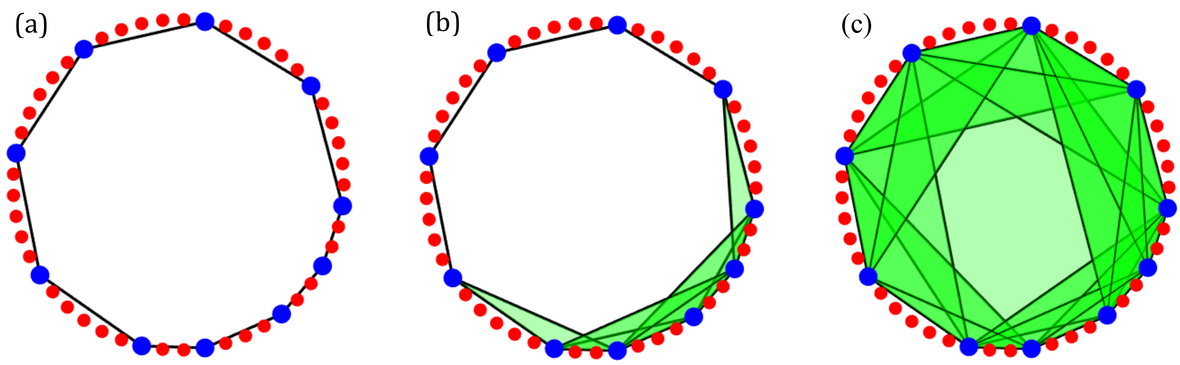

Fig. 9. Lazy Witness Complex for ν = 1: (a) when R = 0.12, (b) when R = 0.69, and (c) when R = 0.72.

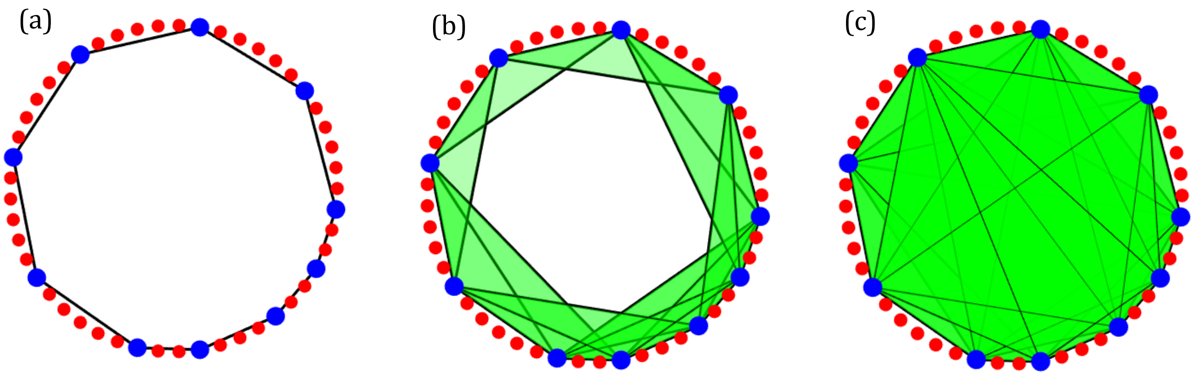

Fig. 10. Lazy Witness Complex for ν = 2: (a) when R = 0.0, (b) when R = 0.61, and (c) when R = 0.64.

The resultant persistence barcodes for $v = 0{,}1$ and 2 have been shown in Fig.11.

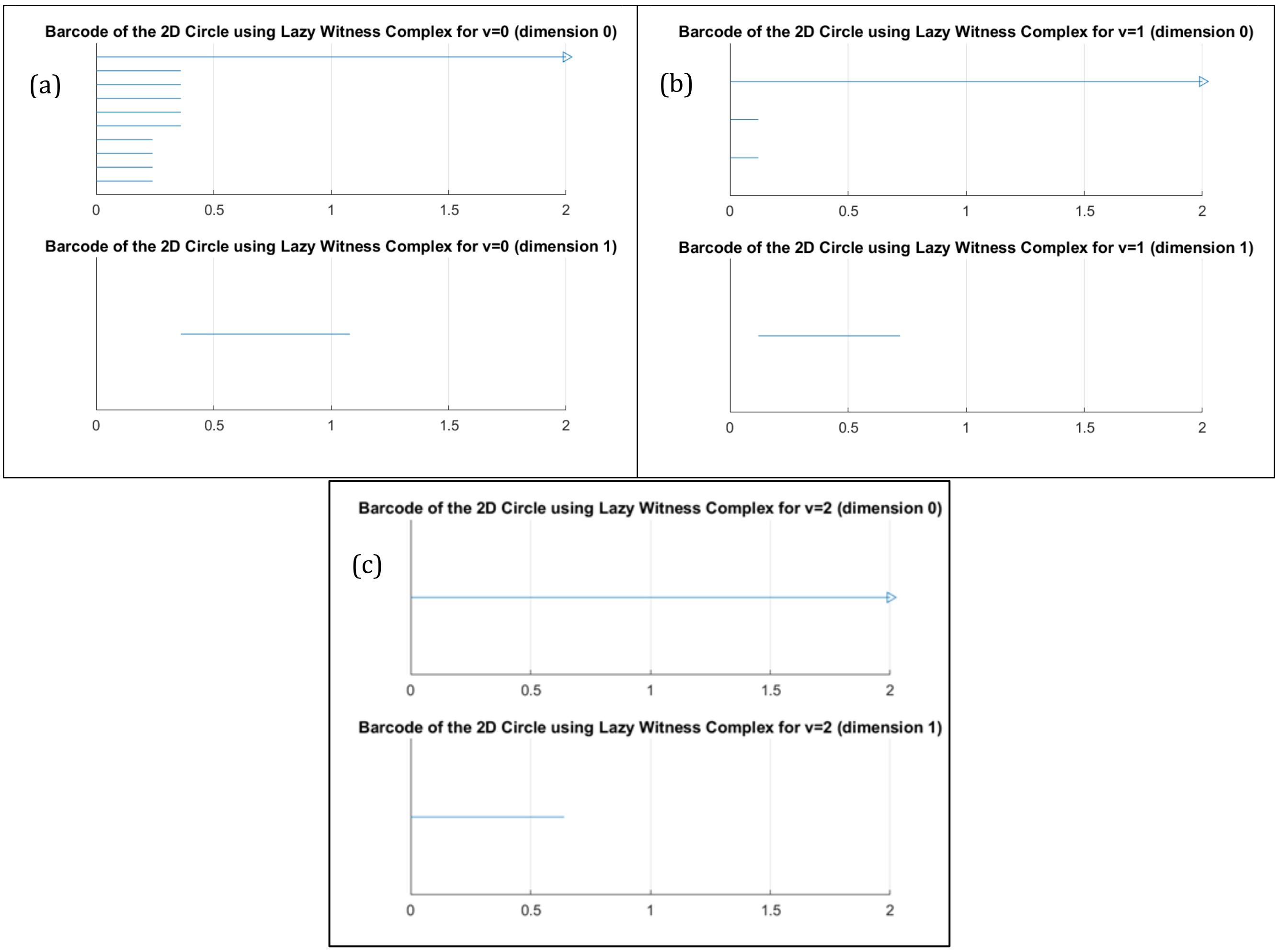

Fig. 11. Persistence barcodes in the case of (a) $v = 0$, (b) $v = 1$ and (c) $v = 2$.

The number of simplices when $v = 0$ is 175. From the persistent barcode for $v = 0$ in Fig. 11(a), we observe that in dimension zero, the persistence intervals are $[0.0, 0.24), [0.0, 0.24), [0.0, 0.24), [0.0, 0.24),$ $[0.0, 0.36), [0.0, 0.36), [0.0, 0.36), [0.0, 0.36), [0.0, 0.36)$, and $[0.0, \infty)$. In dimension one, there is a single persistence interval $[0.36, 1.08)$. As illustrated in Fig. 8, when $R = 0.36$ (see Fig. 8(a)), a loop is born in the complex. This loop persists until $R = 1.08$ (see Fig. 8(c)), indicating its lifespan within this range. Between these values, at intermediate stages such as $R = 1.03$ (see Fig. 8(b)), the witness complex evolves but the loop remains present until it eventually dies at $R = 1.08$. From these observations, we conclude for $v = 0$ that there is one connected component that persists indefinitely and one loop that is born at $R = 0.36$ and dies at $R = 1.08$.

Examining the persistent barcode for $v = 1$ in Fig. 11(b), we find that for dimension zero, the persistence intervals are $[0.0, 0.12)$, $[0.0, 0.12)$, and $[0.0, \infty)$. In dimension one, there is a single persistence interval $[0.12, 0.72)$. As depicted in Fig. 9, a loop emerges in the complex at $R = 0.12$ (see Fig. 9(a)), and it persists until $R = 0.72$ (see Fig. 9(c)), marking its duration. At intermediate stages, such as $R = 0.69$ (illustrated in Fig. 9(b)), the witness complex undergoes transformations, yet the loop remains until it eventually disappears at $R = 0.72$. From this analysis, we conclude that for $v = 1$, the complex has one connected component that persists indefinitely, and one loop that forms at $R = 0.12$ and dissolves at $R = 0.72$. Similarly, a similar pattern occurs for $v = 2$, as shown in Fig. 10. From that figure, we observe that there is one connected component, and a loop is born at $R = 0.00$ at the start of the filtration and disappears at $R = 0.64$.

In Fig. 11, for $v = 0$, we observe that there are more connected components, and they persist for a longer duration, but they are ultimately regarded as noise. For $v = 1$, the amount of noise is less when considering connected components compared to $v = 0$. For $v = 2$, we see that there is no noise, with only one connected component emerging at the very start of the filtration. When examining loops (β_1), we see that for $v = 0$, $v = 1$, and $v = 2$, there is only one loop. However, this loop forms later for $v = 0$ than for $v = 1$ and $v = 2$, with the loop appearing at the beginning of the filtration in the case of $v = 2$. From this example, we can conclude that the case $v = 2$ yields much better results than $v = 0$ or $v = 1$.

6. Discussion

Here, we offer a set of concise observations regarding the three distinct categories of persistent witness complexes corresponding to $v = 0,1,2$:

$\boldsymbol{v = 0}$: The complexes $W(D; R, 0)$ exhibit a strong relationship with the $Rips(L; R)$ complexes. Specifically, there exist the following inclusions:

$$W(D; R, 0) \subseteq Rips(L; 2R) \subseteq W(D; 2R, 0)$$

This inclusion chain allows us to draw significant parallels between the persistent homology groups of the two families. In practical applications, we frequently observe that the interval graphs corresponding to $W(D; R, 0)$ and $Rips(L; R)$ share similar structures, as depicted in Fig. 12. These similarities suggest that, under certain conditions, the $W(D; R, 0)$complex can serve as a computationally efficient alternative to the Rips complex, without sacrificing the essential topological information.

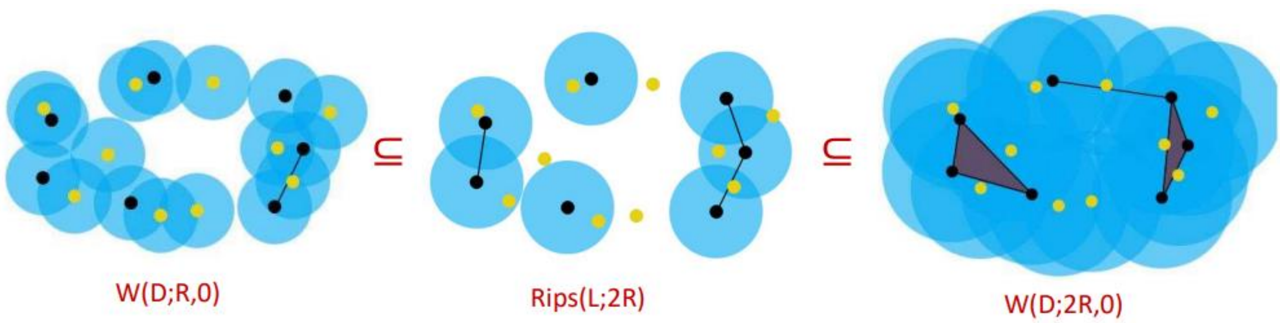

Fig. 12. Relation between Lazy Witness and Rips complex.

$\boldsymbol{v} = \mathbf{1}$: Among the three categories, the $v = 1$ family is perhaps the most theoretically grounded. It can be understood as arising from a sequence of coverings of the space X, where each landmark point is surrounded by a Voronoi-like region. As the radius R increases, these regions begin to overlap, leading to a more connected complex. The conceptual motivation here lies in capturing the gradual merging of local features into a coherent global structure, which is especially useful in applications where the underlying space has a well-defined geometric or topological character. This gradual overlap also helps in controlling the complexity of the resulting complexes, making them more manageable for computation while still preserving critical topological features.

$\boldsymbol{v} = \mathbf{2}$: The rationale behind the $v = 2$ category is less intuitive compared to $v = 1$, yet it leads to an intriguing identity when $R = 0$:

$$W(D; 0,2) = W(D)$$

This identity implies that the complex is essentially accurate at $R = 0$, or requires only a slight increase in R to capture the correct topological information. In practice, the $v = 2$ complexes are observed to produce exceptionally clear persistent interval graphs, with minimal noise. This clarity suggests that the $v = 2$ complexes may be particularly well-suited for applications where reducing noise is critical, such as in data denoising or feature extraction tasks. The robustness of these complexes, even at minimal filtration levels, makes them a valuable tool in topological data analysis, where maintaining a balance between simplicity and accuracy is often challenging.

By understanding these different categories, we can better tailor the choice of witness complex to the specific needs of a given application, whether the goal is to reduce computational complexity, enhance clarity, or preserve topological accuracy. Each category offers unique advantages, making the witness complex a versatile tool in the study of complex data.

7. Conclusion

Extracting information from large-scale data requires significant effort and time due to its unusually large size. In this study, the Lazy Witness Complex is demonstrated, as theory guarantees that selecting a sample of landmarks preserves the topological structure of the data. Families of Lazy Witness Complexes have been computed, and their efficiency has been evaluated based on the outcomes of persistent homology. The methodologies and insights presented here will aid in the continued development and application of TDA techniques, enabling more effective analysis and interpretation of complex datasets.

Moreover, the Lazy Witness complex has drawbacks also. The selection of landmarks can significantly impact the resulting complex. Poorly chosen landmarks may fail to capture important topological features. Also, it has computational complexity. While generally more efficient than full complexes, the Witness Complex can still be computationally intensive for large datasets, especially when determining the witness relationships and the approximation of the original data's topological features depends on the distribution and number of landmarks, potentially leading to less accurate representations. The computational method used in this study can be applied to more complex large-scale datasets to extract topological features, making it useful for feature selection or data classification.

8. Acknowledgment

This study was partially funded by the Ministry of Science and Technology (MoST) of Bangladesh, which awarded fellowships to Mst Zinia Afroz Liza, Md. Al-Imran, and Md. Morshed Bin Shiraj for the 2023–2024 academic year. Tozam Hossain also received research funding from the Research Cell of Bangamata Sheikh

Fojilatunnesa Mujib Science and Technology University. The authors express their gratitude to the MoST and the Research Cell of Bangamata Sheikh Fojilatunnesa Mujib Science and Technology University for their financial support of this research.

9. Conflict of interest

None of the authors are involved in any conflicts of interest.

References

[1] Adams, H., Tausz, A., & Vejdemo-Johansson, M. (2014). JavaPlex: A research software package for persistent (co) homology. In Mathematical Software–ICMS 2014: 4th International Congress, Seoul, South Korea, August 5-9, 2014. Proceedings 4 (pp. 129-136). Springer Berlin Heidelberg.

[2] Adams, H., Emerson, T., Kirby, M., Neville, R., Peterson, C., Shipman, et al. (2017). Persistence images: A stable vector representation of persistent homology. *Journal of Machine Learning Research*, *18*(8), 1-35.

[3] Adams, H. and Segert, J. (2011). Simplicial complex filtration demonstrations in Mathematica. *Software Demonstration*. Retrieved from https://www.math.colostate.edu/adams/research .

[4] Aktas, M. E., Akbas, E., & Fatmaoui, A. E. (2019). Persistence homology of networks: methods and applications. *Applied Network Science*, *4*(1), 1-28.

[5] Alexander, Z., Bradley, E., Meiss, J. D., & Sanderson, N. F. (2015). Simplicial multivalued maps and the witness complex for dynamical analysis of time series. *SIAM Journal on applied dynamical systems*, *14*(3), 1278-1307.

[6] Carrière, M., Chazal, F., Ike, Y., Lacombe, T., Royer, M., & Umeda, Y. (2020). Perslay: A neural network layer for persistence diagrams and new graph topological signatures. In *International Conference on Artificial Intelligence and Statistics* (pp. 2786-2796). PMLR.

[7] Dey, T. K., & Wang, Y. (2022). Computational topology for data analysis. Cambridge University Press.

[8] de Silva, V. (2003). A weak definition of Delaunay triangulation. *arXiv preprint cs/0310031*.

[9] De Silva, V., & Carlsson, G. E. (2004). Topological estimation using witness complexes. In *Proceedings of the First Eurographics conference on Point-Based Graphics (SPBG'04)* (pp. 157-166). Eurographics Association, Goslar, DEU.

[10] Don, A. P., Peters, J. F., Ramanna, S., & Tozzi, A. (2020). Topological view of flows inside the BOLD spontaneous activity of the human brain. *Frontiers in Computational Neuroscience*, *14*, 34.

[11] Don, A. P., Peters, J. F., Ramanna, S., & Tozzi, A. (2021). Quaternionic views of rs-fMRI hierarchical brain activation regions. Discovery of multilevel brain activation region intensities in rs-fMRI video frames. *Chaos, Solitons & Fractals*, *152*, 111351.

[12] Gao, J., Guibas, L. J., Oudot, S. Y., & Wang, Y. (2008). Geodesic Delaunay triangulation and witness complex in the plane. In *Proceedings of the nineteenth annual ACM-SIAM symposium on Discrete algorithms* (pp. 571-580).

[13] Garin, A., & Tauzin, G. (2019). A topological "reading" lesson: Classification of MNIST using TDA. In *2019 18th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 1551-1556). IEEE.

[14] Guibas, L. J., & Oudot, S. Y. (2008). Reconstruction using witness complexes. *Discrete & computational geometry*, *40*, 325-356.

[15] Hamilton, G., Dore, T., & Plumberg, C. (2022). Applications of persistent homology in nuclear collisions. *Physical Review C*, *106*(6), 064912.

[16] Hofer, C. D., Kwitt, R., & Niethammer, M. (2019). Learning representations of persistence barcodes. *Journal of Machine Learning Research*, *20*(126), 1-45.

[17] Kindelan, R., Frías, J., Cerda, M., & Hitschfeld, N. (2021). Classification based on topological data analysis. *arXiv preprint arXiv:2102.03709*.

[18] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

[19] Medbouhi, A. A., Polianskii, V., Varava, A., & Kragic, D. (2023). InvMap and Witness Simplicial Variational Auto-Encoders. *Machine Learning and Knowledge Extraction*, *5*(1), 199-236.

[20] Munkres, J. R. (2018). *Elements of algebraic topology*. CRC press.

[21] Shnier, D., Voineagu, M. A., & Voineagu, I. (2019). Persistent homology analysis of brain transcriptome data in autism. *Journal of the Royal Society Interface*, *16*(158), 20190531.

[22] Zhu, X. (2013). Persistent homology: An introduction and a new text representation for natural language processing. In *IJCAI* (No. 2013, pp. 1953-1959).

[23] Zomorodian, A., & Carlsson, G. (2004). Computing persistent homology. In *Proceedings of the twentieth annual symposium on Computational geometry* (pp. 347-356).