Statistics on the Space of Persistent Diagrams with Applications

By
Waqar Hussain Shah

Supervisor
Dr. Sohail Iqbal

Department of Mathematics
COMSATS University Islamabad
Pakistan

July 24, 2021

Introduction

- Presently in every field of life, we are investigating with a data set that is huge in amount. The technique that we are using is a subfield of Algebraic Topology that is Topological Data Analysis (TDA). TDA offers to explain the geometry and the topological features of the quantitative data.
- PH is a flagship tool of TDA. PH is a homology theory that is used to study the qualitative characteristic of the dataset. It computes a ‘persistent diagram’ and there will be no change in topological features.
- There are many kinds of simplicial complexes such as $\hat{C}$ech, Vietoris-Rips, weak witness, strong witness, and Delaunay complexes. Homology of the simplicial complex gives the algebraic measure based on cycles that are not boundaries.

Introduction

- Also, in computational topology especially in TDA, there are many ways to study logical data. The technique of persistent landscape, Riemannian frameworks, and smooth Euler characteristic transform(SECT).
- In this work, we are dealing with different spaces of Persistent diagrams and applications.

Preliminaries

Category

A category C consists of three ingredients

1. A collection of objects in C, denoted by set obj$C$,
2. Morphisms between objects Mor(θ, ζ) in C, for every ordered pair $\theta, \zeta \in objC$,
3. Composition of morphisms $Mor(\theta, \zeta) \times Mor(\zeta, \eta) \to Mor(\theta, \eta)$ in C, for every $\theta, \zeta, \eta \in objC$.

Such that these three ingredients satisfy the following properties,

1. The family of Mor(θ, ζ)'s is pairwise disjoint,
2. Composition is always associative,
3. For each $\theta \in objC$, there exists an identity $1_\theta \in \mathrm{Mor}(\theta, \theta)$ satisfying

$$1_\theta \circ h = h \qquad h \circ 1_\theta = h.$$

for every morphism h.

Functor

If B and D are categories, a functor $T : B \to D$ is a function, that is,

1. $b \in ObjB$ implies $Tb \in ObjD$,
2. If $f : B \to B'$ is a morphism in B. Then $TF : TB \to TB'$ is a morphism in D,
3. If g, h are morphisms in B for which $g \circ h$ is defined, then

$$T(g \circ h) = (Tg) \circ (Th).$$

Simplex

An affine independent subset let say, $\{u_0,\ u_1, \ldots,\ u_m\}$ of $\mathbb{R}^n$. Convex set spanning by this set, represented by $[u_0,\ u_1, \ldots,\ u_m]$, is called affine m-simplex, and with vertices $u_0,\ u_1, \ldots,\ u_m$.

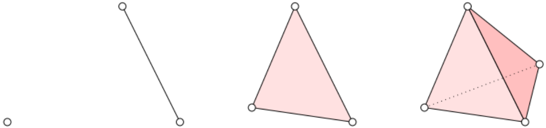
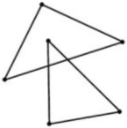


Figure: Simplexes

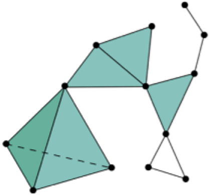
Simplicial Complex

Consider K is a simplicial complex in space of Euclidean and its a collection of finite simplexes such that;

1. If $x \in K$ then each face of s belongs to K.
2. If $x, y \in K$, then $x \cap y$ is either empty or a common face of x and y.



Not a Simplicial Complex



Simplicial Complex

The important algebraic tool that we have
C_qK Free Abelian group with K simplicial complexes having basis Vert(K).

Boundary Operator

Boundary operator is define as $\partial_q : C_qK \to C_{q-1}K$ by setting

$$\partial_q(\langle u_0, u_1, \ldots, u_q \rangle) = \sum_{i=0}^{q} (-1)^i \langle u_0, u_1, , ...\hat{u}_i..., u_q \rangle. \qquad (1)$$

Where $\hat{u}_i$ means delete u_i and extending by linearity.

Homology Groups

If K is an oriented simplicial complex, then

- $Z_q K = \text{Ker}\partial_q$ represents the **qth-simplicial cycles**.
- $B_q K = \text{Im}\partial_{q+1}$ represents the **qth-simplicial boundaries**.
- $H_q K = Z_q K / B_q K$ represents the **Homology group.**

The Fundamental Theorem

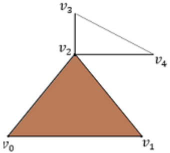
A group G that is Abelian (finitely generated) has the same isomorphism as the direct product of cyclic groups in the form;

$$\mathbb{Z}_{(u_1)^{s_1}} \times \mathbb{Z}_{(u_2)^{s_2}} \times \cdots \times \mathbb{Z}_{(u_n)^{sn}} \times \mathbb{Z} \times \mathbb{Z} \times ...\mathbb{Z},$$

where the u_i are primes, not necessarily distinct, and also in the form

$$\mathbb{Z}_{n_1} \times \mathbb{Z}_{n_2} \times ... \times \mathbb{Z}_{n_s} \times \mathbb{Z} \times \mathbb{Z} \times ... \times \mathbb{Z},$$

where n_i divides n_{i+1}.



$$C_2(K) = \{n\rho : n \in \mathbb{Z}\} \cong \mathbb{Z}.$$

$$C_1(K) = \{m_1\sigma_1 + ... + m_6\sigma_6 : m_i \in \mathbb{Z}\} \cong \mathbb{Z}^6.$$

$$C_0(K) = \{m_1 v_0 + ... + m_5 v_5 : m_i \in \mathbb{Z}\} \cong \mathbb{Z}^5.$$

Betti Numbers β_2(K) = 0, β_1(K) = 1, β_0(K)=1.

Homological Algebra

Exact Sequence

Consider R be a ring with identity, and I, J, K be R- modules. Let

$$f\ :\ I \to J,\ g\ {:}J \to K$$

be R- module homomorphisms. Then the sequence

$$\ldots I \xrightarrow{f} J \xrightarrow{g} K \ldots$$

is exact at J if and only if,

$$(f\ :\ I \to J) = \ (g\ :\ K \to L)$$

If a sequence of modules and homomorphisms is exact at each module in the sequence, then it is said to be exact.

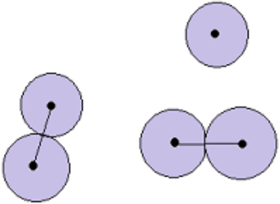
$$\begin{tikzcd}
 & \vdots \arrow[d, "\partial_{i+2}"] & \vdots \arrow[d, "\partial_{i+2}"] & \vdots \arrow[d, "\partial_{i+2}"] & \\
0 \arrow[r] & A_{i+1} \arrow[r, "p_{i+1}"] \arrow[d, "\partial_{i+1}"] & B_{i+1} \arrow[r, "q_{i+1}"] \arrow[d, "\partial_{i+1}"] & C_{i+1} \arrow[r] \arrow[d, "\partial_{i+1}"] & 0 \\
0 \arrow[r] & A_i \arrow[r, "p_i"] \arrow[d, "\partial_i"] & B_i \arrow[r, "q_i"] \arrow[d, "\partial_i"] & C_i \arrow[r] \arrow[d, "\partial_i"] & 0\,. \\
0 \arrow[r] & A_{i-1} \arrow[r, "p_{i-1}"] \arrow[d, "\partial_{i-1}"] & B_{i-1} \arrow[r, "q_{i-1}"] \arrow[d, "\partial_{i-1}"] & C_{i-1} \arrow[r] \arrow[d, "\partial_{i-1}"] & 0 \\
 & \vdots & \vdots & \vdots &
\end{tikzcd}$$

Figure: Exact sequence of chain complexes

The columns of this diagram are chain complexes and rows are exact sequences.

Nerve of a Topological Space

If X is a topological space with covering $\mathcal{P} = \{U_\alpha\}_{\alpha \in I}$, then nerve of $\mathcal{P}$ will be the simplicial complex with the vertex set I, where a k-simplex is spanning by $\{\alpha_1, \alpha_2, \ldots, \alpha_k\}$ if and only if $\{U_{\alpha_0} \cap \cdots \cap U_{\alpha_k}\} \neq 0$.



Complexes

- Čech complex
- Vietoris-Rips complex
- Strong/Weak/Lazy witness complex
- α complex
- Cubical complex
- Clique complex
- CW complex

Vietoris-Rips

If (X, d) be a metric space. Then the vietoris-Rips complex of the space X with some parameter ϵ is represented by $VR(X, \epsilon)$ will be the simplicial complex whose vertex set is X, and where $\{x_0, x_1 \ldots x_k\}$ span a k-simplex iff d$(x_u, x_v) \leq \epsilon \ \forall \ 0 \leq u, v \leq k$.

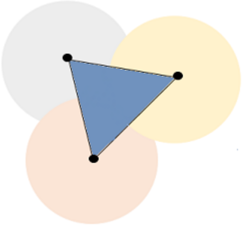


Figure: Vietoris-Rips Complex

Persistent Homology

Persistent homology is a flagship tool of TDA. Persistent homology is a homology theory that is used to study the qualitative characteristic of the dataset. It computes a 'Persistent Diagram' and there will be no change in topological features. In homology, we are interested in the holes of geometrical objects, where the homology groups offer a mathematical language to describes these holes.

- β_0 represent the number of connected components.
- β_1 represent the one-dimensional holes.
- β_2 represent the two-dimensional voids.

Definition

Consider $\mathcal{C}$ be any category, $\mathcal{P}$ a partially ordered set. $\mathcal{P}$ is a category with objects of $\mathcal{P}$, and with distinct morphisms from a to b whenever $a \leq b$. Then $\mathcal{P}$-persistence object in $\mathcal{C}$ refers to a functor $\phi : \mathcal{P} \to \mathcal{C}$.

Chain Complex

A C_* chain complex is a doubly infinite sequence of modules $\{C_i : i \in \mathbb{Z}\}$ over some ring of unity, with homomorphisms $\partial_i : C_i \to C_{i-1}$ for each $i \in \mathbb{Z}$, such that $\partial_i \circ \partial_{i+1} = 0 \ \forall\ i$.

Let C, D be two chain complexes; then a chain map $f : C \to D$ is a collection of morphisms $\{f_n : C_n \to D_n\} \ \forall\ n \in \mathbb{Z}$ such that all the diagrams are commutative,

$$\begin{tikzcd}
C_{n+1} \arrow[r, "g_n"] \arrow[d, "f_{n+1}"'] & C_n \arrow[d, "f_n"] \\
D_{n+1} \arrow[r, "g'_n"'] & D_n
\end{tikzcd}$$

Lifetime of a feature

The period of surviving of these intervals shows the lifetime of a homology group or the homological info by varying ϵ of any simplicial complex construction, enabling us to recover the possible features by overcoming the noise of the data. The intervals are known as barcodes of Persistent homology.

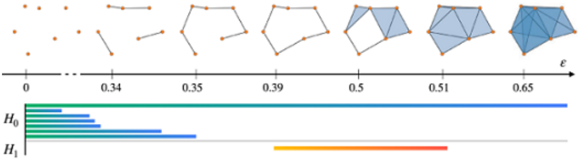
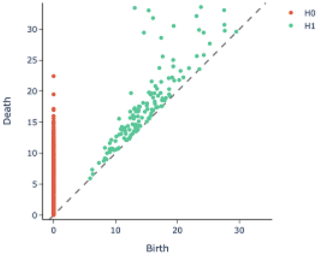


Figure: Feature and its barcode

Persistent Diagram

A persistent diagram (PD) is a multiset that is a union of a finite multiset of points in $\mathbb{R}^2$ with the multiset of points on the diagonal $\triangle = \{(m, n) \in \mathbb{R}^2 | m = n\}$, where an infinite multiplicity of each point on that diagonal.



Persistent Module

Consider k be a field and M be a persistent module of a K-vector spaces $\{M(p)|p \in \mathbb{R}\}$ togeather with K linear map $\{ v_p^q : M(p) \to M(q)|p \leq q\}$ such that,

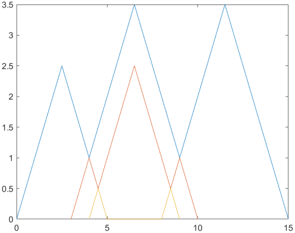
1. $\forall$ p, $v_p^p : M(p) \to M(p)|p \leq q$ is the identity map.
2. if $p \leq q \leq r$ then $v_p^r = v_r^q \circ v_p^q$.

Persistent Landscape

Consider M be a persistent module the persistent landscape is a function $\lambda^* : \mathbb{N} * \mathbb{R} \to \mathbb{R}$ given by,

$$\lambda^*(K, t) = \sup\{h \geq 0 | \mathrm{rank} M(t - h \leq t + h) \geq K\}.$$

The points of a PD are rotated from birth-death pairs (b, d) to $(x, y) = ((d + b)/2, (d - b)/2.$



Properties of Persistent Landscape

1. Invertibility. The mapping is invertible from PDs to persistence landscapes.
2. Stability. Consider D_1 and D_2 be two PDs and also λ_1^* and λ_2^* are their persistent landscape function. Also $\forall$ t and K, $\mid \lambda^*_{1(k)} - \lambda^*_{2(k)} \mid \leq d_B(D_1, D_2)$, where d_B denotes the bottleneck distance.
3. parameter. There is no role of parameters in persistent landscape.
4. Nonlinearity and Computability of persistence landscapes. If the $D1$ and $D2$ are two persistent diagrams and S is the linear vector of PDs then $S(D1 \cup D2) = S(D1) + S(D2)$.

Persistence Images

Two standard ways to represent PH information are PD and barcodes. These tools indicate at which scale (parameters) topological features first appear are ‘born’ and no longer remain ‘die’. There is a barrier that how we can use machine learning tasks based on PDs in a parallel way. In this era, there is still not a concrete answer to when and how to use machine learning and computational topology at the same time. The solution of this fundamental problem of a representation of PDs is a Persistence images.

Definition

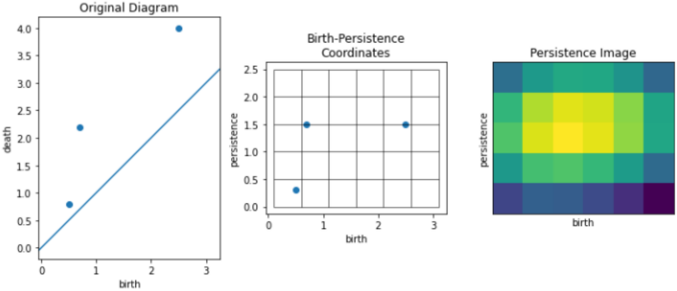
A Persistence image (PI) a finite dimensional vector representation of a PD.

Framework of PIs

Definition

A map of PDs, D to an integrable function $\omega_D : \mathbb{R}^2 \rightarrow \mathbb{R}$, is known as a persistence surface.

First, we map PD to an integral function called persistence surface. The stability surface ω is determined as the sum of weighted Gaussian functions centered at every point in the PD. Then, a discretization is made by the stability surface sub-domain which outputs in a mesh. As a result, PI is acquired by integrating the stability surface over every mesh square, which gives us a pixel value matrix.


Original Diagram
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
death
0
1
2
3
birth
Birth-Persistence
Coordinates
2.5
2.0
1.5
1.0
0.5
0.0
persistence
0
1
2
3
birth
Persistence Image
persistence
birth

Joint Probability Density Function

Let Y_1, Y_2 be jointly continuous random variables, if there exists a positive function

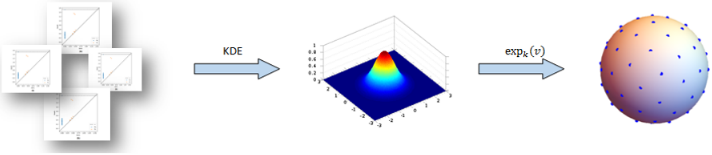
$$g_{Y_1 Y_2} : \mathbb{R}^2 \to \mathbb{R}$$

where any set $L \in \mathbb{R}^2$ such that $(Y_1, Y_2) \in L$ we have,

$$\mathcal{P}(L) = \int \int_L g_{Y_1 Y_2}(y_1, y_2) dy_1 dy_2. \tag{2}$$

the function $g_{Y_1 Y_2}(y_1, y_2)$ here represents a joint probability density function of y_1 and y_2.

In Riemannian framework persistent diagrams are approximated as 2D probability density functions by applying kernel density estimation, with a Guassian kernal of variance σ^2 and mean zero.



KDE

$\exp_k(v)$

Smooth Euler Characteristics Transform

Persistent Homology Transform (PHT) is a statistical tool to achieve statistical shape interpretation on objects and shapes in $\mathbb{R}^3$ and $\mathbb{R}^2$.

Definition

The smooth Euler characteristic curve (SEC), for a fixed direction $v \in S^{d-1}$ is describe as, $\forall\ n \in \mathbb{R}$ then,

$$\text{SEC(K)} : \mathbb{R} \rightarrow \mathbb{L}^2.$$
$$F_v^K(n) = \int_{-\infty}^{n} Z_v^K(m) dm.$$

Euler Characteristic

Definition

The topological space S, $H_k(X)$ denote the k-th homology group of S, and β_k denote the homology group's rank. The alternating sum of S is the Euler Characteristic(EC) $\chi(S)$.

$$\chi(S) = \beta_0 + \beta_1 + \beta_2 \cdots = \textstyle\sum_{k=0}^{\infty}(-1)^k \beta_k.$$

Same as the EC, for a discrete shape or three-dimensional surface, may be described by number of K simplices as a simplicial complex K;

$$\chi(S) = V - E + F.$$

Definition

The smooth Euler characteristic transform (SECT) of a shape $M \subset \mathbb{R}^d$ for a simplicial complex K, with $d = \{2{,}3\}$, is the map

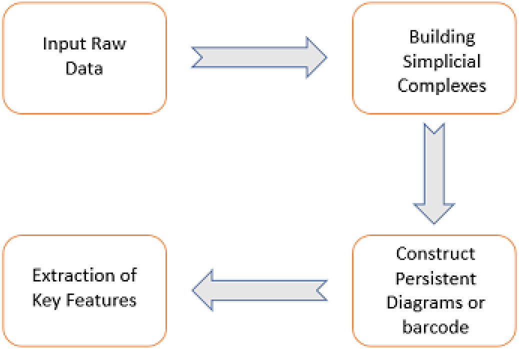
$$\mathrm{SECT(K)} : S^{d-1} \to \mathbb{L}^2[a_v, b_v].$$
$$v \to F_v^K(b_v)$$

for all $v \in S^{d-1}$. Each curve F_v^K is also lies in the space of Hilbert $\mathbb{L}^2$.

Computational Techniques For PH

1. JavaPlex
2. Ripser
3. Dionysus
4. Perseus
5. PHAT
6. GUDHI
7. DIPHA
8. Persim

Computational steps for PH



Input Raw Data

Building Simplicial Complexes

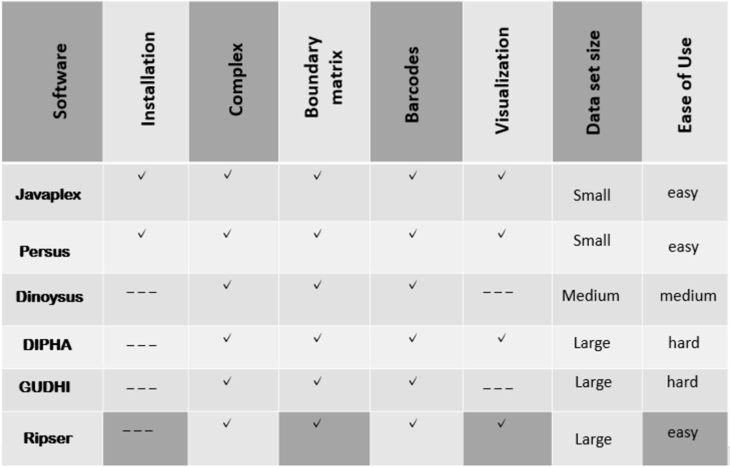
Construct Persistent Diagrams or barcode

Extraction of Key Features

Package	**time in seconds**
Dionysus	–
Ripser	2
GUDHI	381
DIPHA	926
JavaPlex	13,607
Perseus	–

Table: Elapsed time in seconds for each package

Limitation of Packages



Software	Installation	Complex	Boundary matrix	Barcodes	Visualization	Data set size	Ease of Use
Javaplex	✓	✓	✓	✓	✓	Small	easy
Persus	✓	✓	✓	✓	✓	Small	easy
Dinoysus	– – –	✓	✓	✓	– – –	Medium	medium
DIPHA	– – –	✓	✓	✓	✓	Large	hard
GUDHI	– – –	✓	✓	✓	– – –	Large	hard
Ripser	– – –	✓	✓	✓	✓	Large	easy

Basic Hematology

Blood is a body fluid that flows in the blood vessels of all animals. RBCs (red blood cells), WBCs (white blood cells), and Platelets are the three primary components of blood.

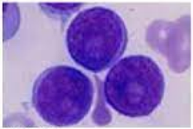
What is ALL?

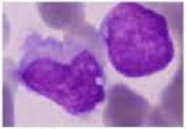
Acute Lymphocytic Leukemia (ALL) which is also called Acute Lymphoblastic Leukemia is a type of pervasive childhood blood cancer that occurs with rapid and continuous production of WBCs and after all, it disturbs the immune system.

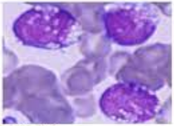
FAB Classification

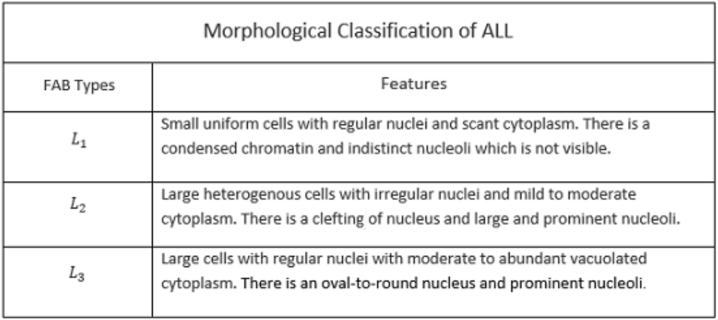
A number of classifications of hematological diseases are defined in French American British (FAB) classification systems. It was released for the first time in 1976. ALL is divided into three subtypes under the FAB categorization system. There is a big challenge between the classification because both normal and ALL cells are morphological same.

FAB Classification


ALL-L1


ALL-L2

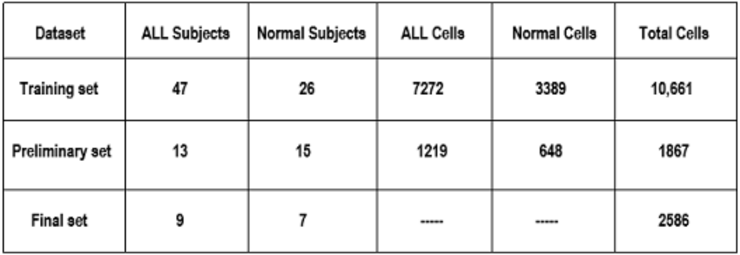

ALL-L3



Morphological Classification of ALL	
FAB Types	Features
L_1	Small uniform cells with regular nuclei and scant cytoplasm. There is a condensed chromatin and indistinct nucleoli which is not visible.
L_2	Large heterogenous cells with irregular nuclei and mild to moderate cytoplasm. There is a clefting of nucleus and large and prominent nucleoli.
L_3	Large cells with regular nuclei with moderate to abundant vacuolated cytoplasm. There is an oval-to-round nucleus and prominent nucleoli.

Dataset of C-NMC-2019

The Cancer Imaging Archive (TCIA) has made the C-NMC-2019 dataset. This C-NMC-2019 dataset was also used in the medical imaging challenge Classification of Normal vs Malignant Cells in B-ALL White Blood Cancer Microscopic Image: ISBI 2019.



Dataset	ALL Subjects	Normal Subjects	ALL Cells	Normal Cells	Total Cells
Training set	47	26	7272	3389	10,661
Preliminary set	13	15	1219	648	1867
Final set	9	7	-----	-----	2586

Image segmentation

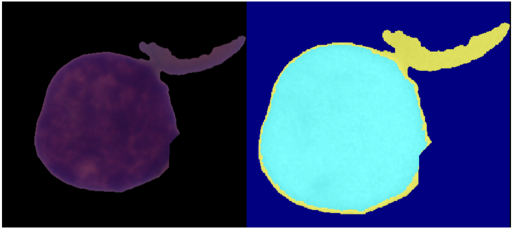


Figure: Original vs Segmented Image

Image segmentation

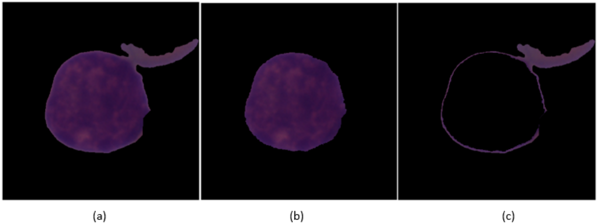
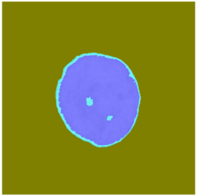
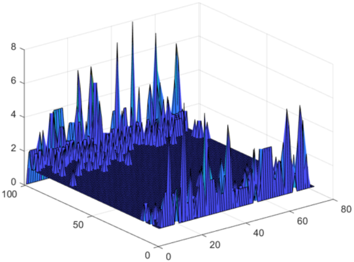


Figure: In second there is a Nucleus and in third there is a Cytoplasm of an image

Results of smooth Euler characteristics transform

In FAB classification its given that the nucleus of L_1 and L_3 is uniform while in L_2 there is a clefting in nucleus, so SECT determines the euler curves (clefting) of the images.





Machine Learning

Supervised Machine Learning

Data in the training sample containing information on the available inputs and their labelled outcomes for some certain behaviour. This approach is said to supervised ML.

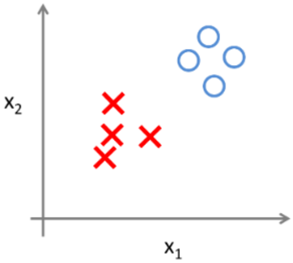


Figure: Clearly labeled data as circles and crosses

Machine Learning

Supervised Machine Learning

Unsupervised ML, does not categorized incoming data with specific labels; instead, the machine generates response based on similarities between the input data.

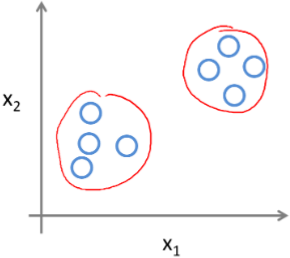


Figure: The clustering are being formed from unlabeled data

Machine Learning

Support Vector Machine

SVM is a supervised machine learning method that learns by dividing data into categories or labels. In a support vector machine, data is separated using hyperplanes. For example, if the data is on a 2D plane, the hyperplane that separates data sets for prediction is a line.

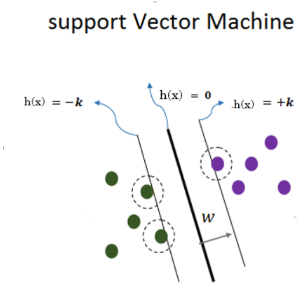


Figure: SVM separating the data of blue and red dots

Results

Results	
Method	F1 Score
Our Model for Cancer	78%

Table: Results for Cancer

Results	
Method	F1 Score
Our Model for Leukemia cells	78%
Pan	91.0%
Xia	84.8%
Ding	85.5%
Gehlot	90.4%

Table: Results for Leukemia cells

Conclusion and Future Work

In this work, we have studied the TDA techniques and their current implementations in the different fields of research. In particular, we have seen the TDA tools, Persistent Homology as a discriminatory technique among the images of Normal and ALL (Acute Lymphoblastic Leukemia) cells. Moreover, we have used Machine Learning tools for classification. This effort will lead to significant progress in the field of medical imaging. We have applied the image analysis techniques on microscopic images for their classification as pathologists can do.

Conclusion and Future Work

In the future, we are plan to build a model to detect the effective topological descriptor for the best and accurate FAB classifications that depend upon geometrical features. Our aim is this work will be able to be carried out in different medical imaging issues and better results can be achieved by utilizing the different metrics on the space of Persistent diagrams.

Refrences

- Carlsson, Gunnar. "Topology and data." Bulletin of the American Mathematical Society 46.2 (2009): 255-308.
- Otter, Nina, et al. "A roadmap for the computation of persistent homology." EPJ Data Science 6.1 (2017): 17.
- Bubenik, Peter, and Paweł Dłotko. "A persistence landscapes toolbox for topological statistics." Journal of Symbolic Computation 78 (2017): 91-114.
- Anirudh, Rushil, et al. "A Riemannian framework for statistical analysis of topological persistence diagrams." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2016.
- Crawford, Lorin, et al. "Predicting clinical outcomes in glioblastoma: an application of topological and functional data analysis." Journal of the American Statistical Association 115.531 (2020): 1139-1150.

- Gehlot, Shiv, Anubha Gupta, and Ritu Gupta. "SDCT-AuxNet: DCT augmented stain deconvolutional CNN with auxiliary classifier for cancer diagnosis." Medical image analysis 61 (2020): 101661.
- Bodzas, Alexandra, Pavel Kodytek, and Jan Zidek. "Automated detection of acute lymphoblastic leukemia from microscopic images based on human visual perception." Frontiers in Bioengineering and Biotechnology 8 (2020): 1005.
- Rotman, Joseph J. An introduction to algebraic topology. Vol. 119. Springer Science Business Media, 2013.
- Mileyko, Yuriy, Sayan Mukherjee, and John Harer. "Probability measures on the space of persistence diagrams." Inverse Problems 27.12 (2011): 124007.

Thank you!