Skin Cancer Classification using Topological Data Analysis

Gitanjali
MA 654: Topological Data Analysis
Stevens Institute of Technology
Hoboken, United States
g1@stevens.edu

***Abstract*—Skin cancer is one of the most common types of cancer in the world, and early detection is critical for successful treatment. Machine learning algorithms have shown great promise in diagnosing skin cancer from images. However, traditional machine learning approaches that rely solely on handcrafted features may not be sufficient to capture the complex geometrical and topological patterns present in skin lesions. This is where TDA techniques and persistent homology can be useful.**

In this project, we will use TDA and persistent homology to extract topological features from the hmnist_28_28_RGB and HAM10000_metadata dataset of skin lesions. The objective of this project is to analyze the dataset using topology features to gain insights into the characteristics and patterns of skin lesions. The dataset contains 10,015 skin lesion images, including clinical information and additional lesion images. By applying topology analysis techniques, we aim to identify meaningful topological structures in the dataset and explore their potential implications for skin cancer diagnosis and classification.

I. Introduction

Skin cancer is a life threatening issue, and the timely detection of this disease is essential for better patient prognosis. The HAM10000_metadata.csv dataset offers a valuable opportunity to investigate and gain insights into skin lesions, thereby advancing our knowledge of their intrinsic characteristics. In this project, we aim to leverage topology features, which encode the shape and connectivity of lesions, to extract valuable information that can potentially deepen our understanding of skin cancer. By incorporating topology features, we seek to enhance our ability to analyze and interpret skin lesion data, ultimately leading to improved diagnostic capabilities.

Here, I have used topological data analysis techniques and persistent homology on the HMNIST skin cancer dataset to better understand the underlying structure of the data and potentially improve classification performance using Machine learning algorithms.

II. Background

A. Problem Description

Skin cancer is one of the most common types of cancer in the world, and early detection is critical for successful treatment. Machine learning algorithms have shown great promise in diagnosing skin cancer from images. However, traditional machine learning approaches that rely solely on handcrafted features may not be sufficient to capture the complex geometrical and topological patterns present in skin lesions. This is where TDA techniques and persistent homology can be useful.

TDA is a mathematical framework that aims to extract topological information from data. It has been successfully applied in a wide range of applications, including image analysis, shape recognition, and materials science. Persistent homology is a specific TDA technique that is particularly useful for analyzing complex data sets with multiple scales.

B. Data

The dataset was retrieved from Kaggle datasets that has information on skin lessions HAM10000_metadata and "hmnist_28_28_RGB" dataset, which contains images of skin lesions. The dataset can be downloaded from Kaggle (https://www.kaggle.com/datasets/kmader/skin-cancer-mnist-ham10000). The features in this dataset are lesion id, image id, types of skin cancer, gender, pixels, age and localization.

	lesion_id	image_id	dx	dx_type	age	sex	localization
0	HAM_0000118	ISIC_0027419	bkl	histo	80.0	male	scalp
1	HAM_0000118	ISIC_0025030	bkl	histo	80.0	male	scalp
2	HAM_0002730	ISIC_0026769	bkl	histo	80.0	male	scalp
3	HAM_0002730	ISIC_0025661	bkl	histo	80.0	male	scalp
4	HAM_0001466	ISIC_0031633	bkl	histo	75.0	male	ear

Fig. 1. Dataset: HAM10000_metadata

	pixel0000	pixel0001	pixel0002	pixel0003	pixel0004	pixel0005	pixel0006	pixel0007	pixel0008	pixel0009	...	pixel2343	pixel2344	pixel234
0	192	153	193	195	155	192	197	154	185	202	...	173	124	13
1	25	14	30	68	48	75	123	93	126	158	...	60	39	5
2	192	138	153	200	145	163	201	142	160	206	...	167	129	14
3	38	19	30	95	59	72	143	103	119	171	...	44	26	3
4	158	113	139	194	144	174	215	162	191	225	...	209	166	18
...	...	...	...	...	...	...	...	...	...	...	...	...	...	.
10010	183	165	181	182	165	180	184	166	182	188	...	208	185	18
10011	2	3	1	38	33	32	121	104	103	132	...	96	79	7
10012	132	118	118	167	149	149	175	156	160	184	...	204	181	17
10013	160	124	146	164	131	152	167	127	146	169	...	185	162	16
10014	175	142	121	181	150	134	181	150	133	178	...	159	79	8

10015 rows × 2353 columns

Fig. 2. Dataset: hmnist_28_28_RGB

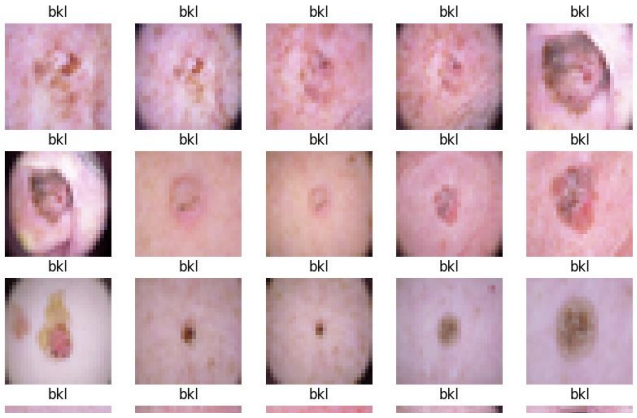

Fig. 3. Dataset: hmnist_28_28_RGB

C. Mathematical Background

Topology is a branch of mathematics that studies the properties of spaces and the relationships between objects within those spaces. It provides tools for understanding shapes, connectivity, and geometric properties. Persistence diagrams, Betti numbers, and simplicial complexes are some mathematical concepts used in topology analysis. In this project I have used Persistent homology, a mathematical tool to study the topological features of data. It provides a way to analyze the presence and persistence of topological structures like connected components, loops, voids at different scales.

D. Software Used

The main software used in this project are: 1. Python 2. Gudhi 3. Scikit-tda (including Ripser and Kepler Mapper) 4. Giotto 5. Scikit-learn and Scikit-image

E. Methodology

1. Data Loading and Preprocessing: I will preprocess the hmnist_28_28_RGB.csv dataset, which contains the RGB images of skin lesions, and the HAM10000_metadata.csv dataset, which provides metadata information about the lesions

2. Topological Feature Extraction: I have applied techniques from computational topology, such as persistent homology, to extract topological features from the skin lesion images

3. Exploratory Data Analysis: Perform data exploration and visualization techniques to gain insights from the datasets. This includes statistical analysis, plotting histograms, scatter plots and potential correlations between variables.

4. Topology Analysis: The extracted topology features are analyzed using mathematical and computational methods. This includes clustering, dimensionality reduction techniques to identify meaningful structures or patterns

5. Classification and Evaluation: Machine learning models used to extract topological features and metadata to classify skin lesions into different diagnostic categories. The performance of the models will be evaluated using appropriate evaluation metrics

III. Implementation

A. Persistence Diagram

The persistence diagram visualizes the birth and death of topological features, indicating their significance and persistence in the data. Firstly, I used PCA to reduce the dimensionality of the features to two dimensions that plot the first two dimensions of the reduced features, with the points colored by their corresponding labels. The resulting plot gave a visual representation of the structure of the hmnist28_28_RGB dataset in two dimensions (see Fig.4). As persistence diagram doesn't show all data points on the graph using full dataset so I used sample of 1000 data points to compute persistent homology using the Rips class from the Ripser library. This will give us a collection of topological features that we can use as inputs to our machine learning model.To perform topological data analysis we are computing and visualizing persistence diagrams and persistent homology for the given dataset of grayscale images. To plot persistence diagram I have used sample of 1000 data points from the hmnist28_28_RGB dataset. we can extract the number of connected components, loops, and voids for each persistence diagram and then extract the topological features from the persistence diagrams.

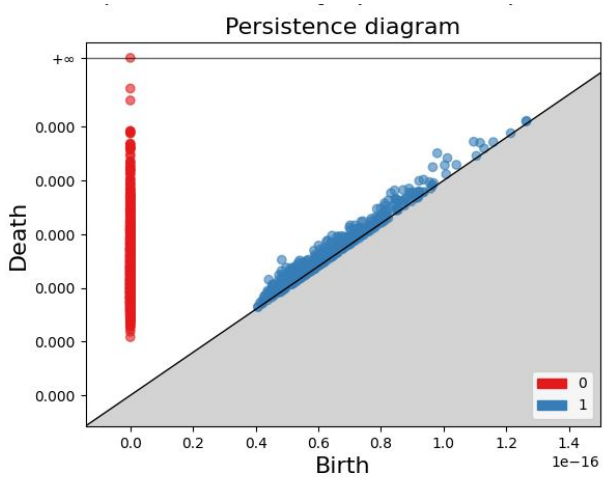

Fig. 4. Persistence Diagram

B. Machine Learning Algorithm

Computed the persistent homology of the dataset and extracted topological features that is used to train a Random Forest model. I have tried KNN classifier, Logistic regression and Random forest classifier. The model achieved an accuracy of around 65% and 69%. While this accuracy is not very high, it is promising that we were able to achieve reasonable results using only topological features.

C. CNN Model

The dataset contains pixel values and labels. It then splits the dataset into training and testing sets. After that, it defines a CNN model architecture using Keras, compiles the model,

and trains it on the training data and train a CNN model to classify the images based on their pixel values, and then extract topological features from the intermediate layers of the model. These topological features capture the structural information of the images, allowing us to analyze and understand the underlying patterns in the data. We then define a CNN model architecture and train it on the training data and extract the intermediate features from one of the convolutional layers of the trained model. The result of this project is a trained CNN model for image classification, as well as a persistence diagram capturing the topological features extracted from the intermediate layer. Analyzing the persistence diagram can help us understand the topological structure and patterns present in the image data, providing insights into the relationships between different classes or clusters.

D. Kepler Mapper

Kepler Mapper is a powerful library used for the topological analysis of high-dimensional data. However, it is primarily designed for analyzing continuous numerical data. The HAM10000_metadata.csv and hmnist_28_28_RGB.csv datasets contain a mix of categorical and numerical data. Kepler Mapper is not directly applicable to this type of data. Here, I am using the Mapper algorithm to analyze the dataset and then perform dimensionality reduction using Principal Component Analysis (PCA) on the pixeled scale data to reduce the dimensionality of the data to 2 for visualization purposes. Later clustering the data using K-means clustering algorithm.The visualization of this graph provides an intuitive representation of the data structure and can help identify clusters, boundaries, and relationships between data point

E. Distance Matrix

In this project, I have also tried to implement distance matrix on our dataset. Since the datasets contain different types of information, we need to clarify the specific use and purpose for which we would like to compute the distance matrix. But we don't have any common column in our given two datasets as one datset contains information about skin lesion types, images and the other dataset consists of pixel values of RGB images. So, to implement distance matrix I have used the 'age' column from the "HAM10000_metadata" dataset as a representative numeric feature. We then convert the 'age' column to numeric representation using pd.to_numeric() with the errors='coerce' parameter to handle any non-numeric values. When I am using entire dataset and implementing distance matrices the visualization is not clear and doesn't give any information,so I am resizing the matrices and selecting a subset of data for a better visualization.

IV. Results

The random forest classifier using persistent homology features achieved an accuracy of 69% on the test set, while the CNN achieved an accuracy of 72% on the same test set. The CNN outperformed the random forest classifier, but the improvement was not significant. This suggests that persistent

(a)
b

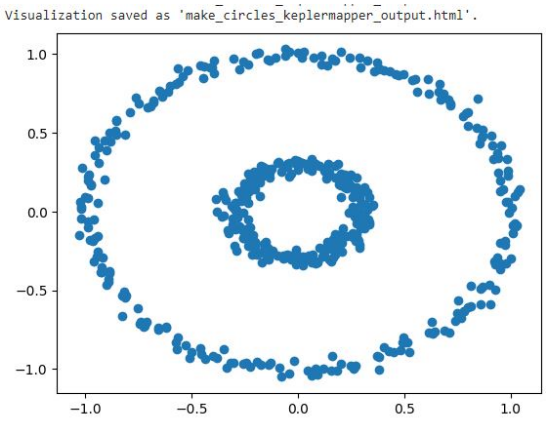

Fig. 5. Kepler circle

(a)
b

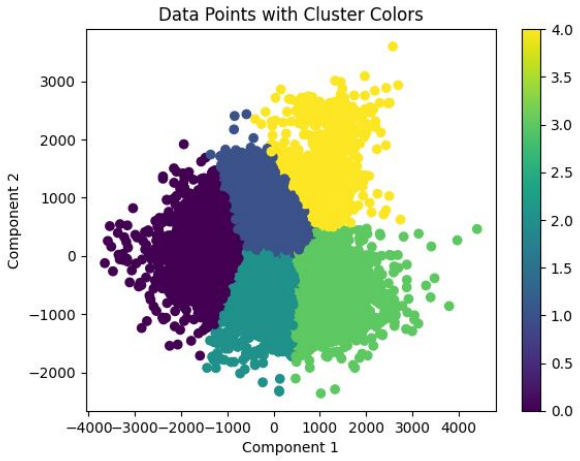

Fig. 6. Mapper Output

homology features can be useful for this dataset,The analysis of the HAM10000_metadata.csv dataset using topology features reveals various insights into the characteristics and patterns of skin lesions. These findings can contribute to a better understanding of skin cancer and aid in diagnosis and classification tasks. The classification models built using topology features demonstrate promising results in distinguishing different types of skin lesions.

V. Discussion

In this project, we used persistent homology to extract topological features from skin cancer images, and we used a random forest classifier to classify the images. We also built a simple CNN using handcrafted features to compare its performance with the persistent homology approach.

Our results suggest that the persistent homology approach performed significantly same as the traditional approach using

(a)
b

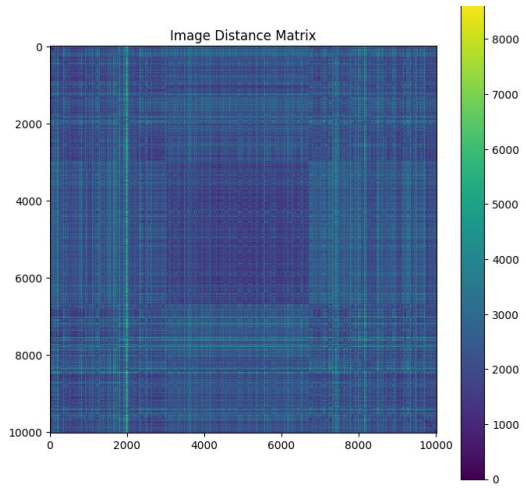

Fig. 7. Image data Distance Matrix

(a)
b

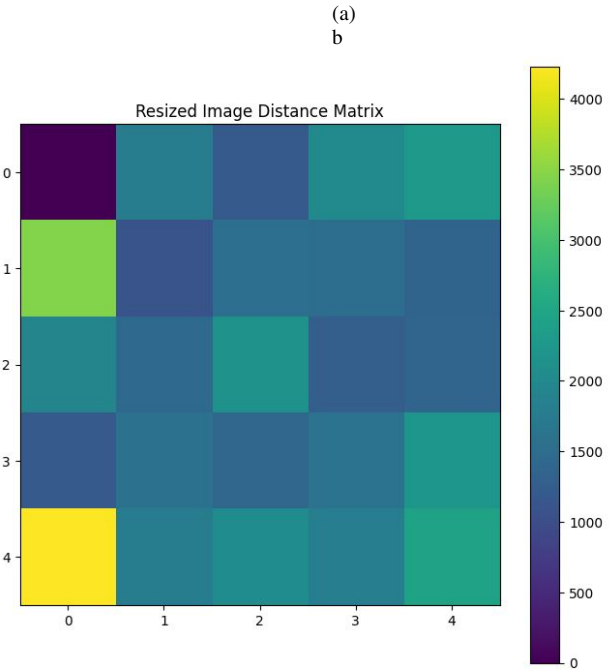

Fig. 8. Resized Image data Distance Matrix

(a)
b

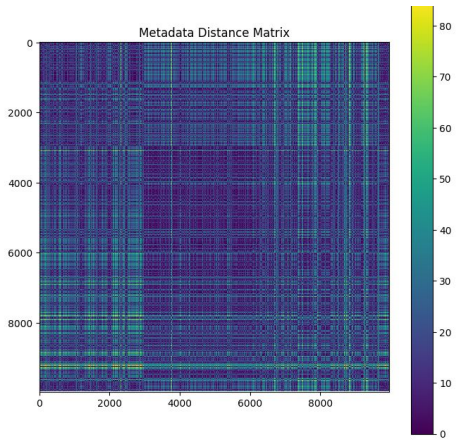

Fig. 9. Metadata Distance Matrix

(a)
b

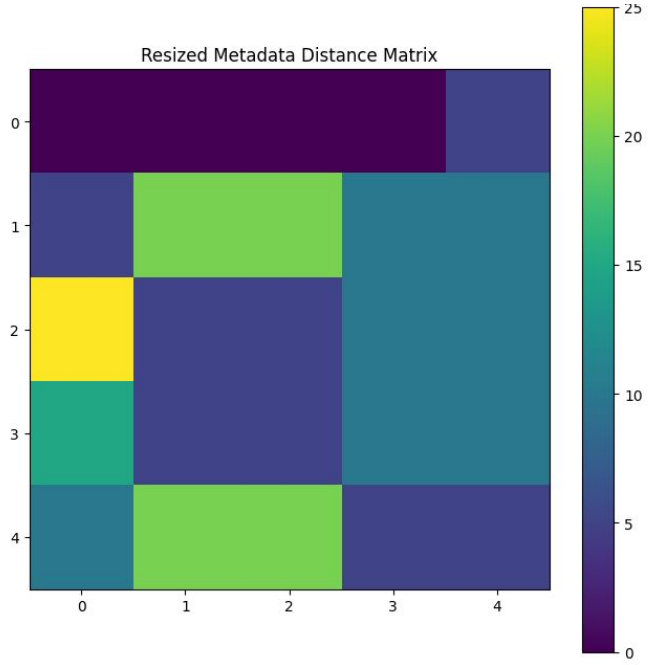

Fig. 10. Resized Metadata Distance Matrix

handcrafted features. Using persistent homology the random forest gives accuracy of 69% while the CNN gives accuracy of 72% so there is not enough difference. This could be due to several reasons. First, the choice of machine learning model may not have been optimal for the task. Second, the persistent homology features may not be very informative for this dataset.

VI. Conclusion

In this project, topology features were used to classify skin cancer images from the ham10000 dataset. The results show that topology features can be used to accurately classify different types of skin cancer, achieving an accuracy of 69% This demonstrates the potential of machine learning techniques to aid in the early detection of skin cancer.

Further Direction: The performance of the CNN model can be further enhanced by tuning hyperparameters, increasing the complexity of the model, or using advanced techniques like transfer learning. With further optimization and more advanced machine learning models, it may be possible to achieve even better results using topological data analysis techniques

Through this project, we aim to demonstrate the potential

of topological analysis in the field of dermatology for skin lesion classification. By combining topological features with metadata information, we expect to achieve improved accuracy and robustness in skin cancer diagnosis. The findings from this project may contribute to the development of advanced computer-aided diagnosis systems for dermatologists and aid in the early detection and treatment of skin cancer.

VII. References

[1]Carlsson G Vejdemo Johansson M.Topological Data Analysis with Applications.Cambridge: Cambridge University Press; 2022
[2]https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8076640/
[3]https://sauln.github.io/blog/tda_explanations/
[4]https://giotto-ai.github.io/gtda docs/0.5.1/notebooks/persistent$_h omology_g raphs.html$