A Topology-Based Machine Learning Framework for Breast Cancer Subtype Classification in Histology Images

Ryan Huang[1,2], Lorin Crawford[3,4]

Program in Liberal Medical Education, Brown University[1], Department of Computer Science, Brown University[2], Department of Biostatistics, Brown University[3], Microsoft Research New England[4]

Abstract

Problem: Subtyping breast cancer is currently a labor-intensive and time-consuming process, relying on immunohistochemical analyses of cell samples, which demand expertise.
Aim: Our goal is to create a computational approach that can accurately differentiate breast cancer subtypes. Currently, state-of-the-art approaches utilize shape-based classification algorithms; however, due to variations found in IDC histologies, they are fairly inaccurate.
Hypothesis: Here, we hypothesize that using a two-step machine learning framework will overcome this interclass heterogeneity.
Approach:
1. First, we will use topological data analysis (TDA) to create a quantitative summary statistic of each histology image to encode their geometry and topology.
2. Afterward, we will parse these values in a machine learning model (Random Forrest) to classify the images.

Conclusion: Overall, this study will provide an interpretable classification framework that can help differentiate breast cancer subtypes, yielding potential application to classify and analyze medical images universally.

Introduction

Current Solutions:
- Using deep learning after segmenting histology image into pieces[1]
  - Results in loss of information
- Multi-omic approaches (molecular and genetic data with a classifier)[2]
  - Needs too much information, which requires more resources

Why Topology and How it Works?
- Subtypes are known to have architectural differences in their histological organization[3]
  - We can extract these differences by analyzing their topology
- We create persistence diagrams, which calculate the birth of vertices, loops, and voids that form from expanding a "ball" from each point in the point cloud
  - The birth and death of these various dimensions corresponds to points in the diagram (shown in the schematic step (iv))

Methods

Intuition of My Approach:
- First, we want to render each image as a condensed point cloud
  - Get coordinates of all the nuclei (using stardist)
  - Use agglomerative clustering (reduces overall run time) and preserves general topology/shape of image[4]
- Then, we can make persistence diagrams from this point cloud and turn them into landscapes for my classifier to easily classify the landscapes.

Data Source:
- Data was taken from: NIH National Cancer Institute GDC Data Portal[5]
- Breakdown of data (504 images):
  - 87 Basal
  - 40 Her2
  - 313 Luminal A
  - 94 Luminal B

NIH NATIONAL CANCER INSTITUTE GDC Data Portal

Framework Schematic

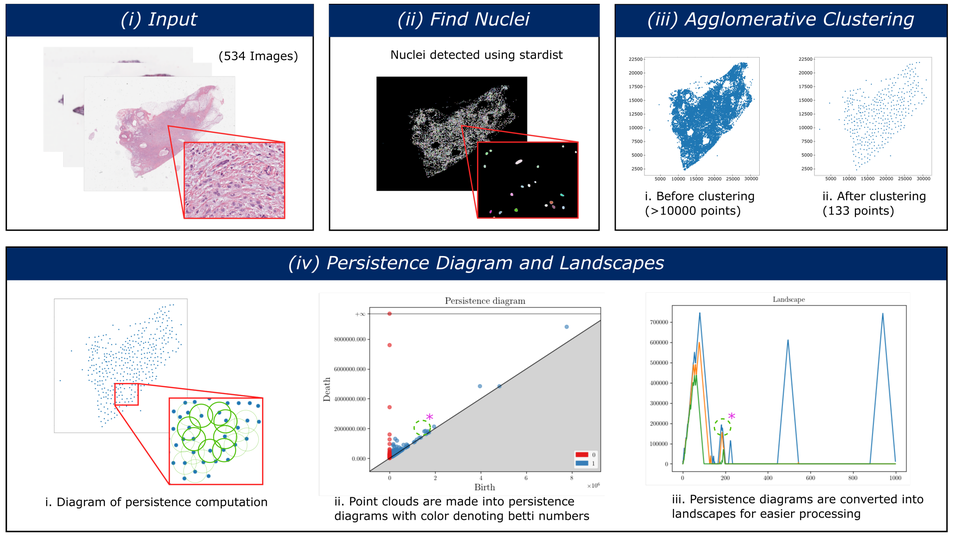
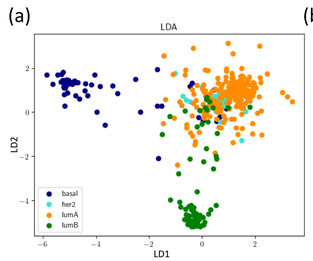


Figure 1. Framework Schematic. The schematic showcases how each image was transformed into a landscape; schematic shows example for one image. (i) This step entailed parsing the images from the database to an interpretable svs file. (ii) After retrieving all the images, we used Stardist, an open-source software, to detect all the coordinates of the nuclei using object detection. (iii) The initial point cloud has over 10000 points, which would be too much for a computer to handle. By employing agglomerative clustering, we reduce the number of points significantly while preserving general shapes and proportions as shown above. (iv) Finally, the condensed point clouds are made into persistence diagrams and, subsequently, into landscapes as shown.

Key Results



(b)

Compared Subtypes		Average Accuracy
Basal	Her2	68.0%
Basal	Luminal A	72.5%
Basal	Luminal B	63.8%
Her2	Luminal A	47.9%
Her2	Luminal B	59.7%
Luminal A	Luminal B	72.3%

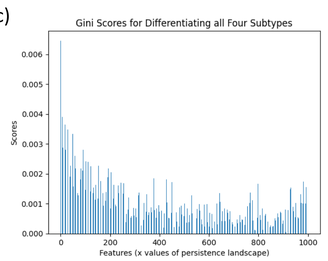


Figure 2. Results from Classification of Landscapes. (a) Linear Discriminant Analysis (LDA) was performed on the landscapes generated from the images. It was clear that Basal and Luminal B subtypes had the greatest difference, while Luminal A and Her2 subtypes were fairly similar. (b) A random forest classifier was used to differentiate the subtypes, and it appears that Basal and Luminal A had the greatest differentiating accuracy, and Her2 and Luminal A had the lowest (as expected from the LDA analysis). (c) Gini scores were computed for each feature, where every feature was the 1000 points in the persistence landscape. This was computed to find which part of the diagram was most important for differentiating all four subtypes overall.

Conclusions

- Agglomerative clustering significantly reduces run times of computation while maintaining the image's shape.
- Using landscapes to summarize a histology image helps preserve information of the original image and can decode the overall architecture and arrangement of the nuclei
- My classifier can help differentiate luminal A breast cancer with basal and luminal B with fairly decent accuracy. More data might be needed for Her2 to increase the classification accuracies.

Future Directions

- (Currently in-progress) Backtrack from gini scores and project these scores to the point cloud. Then, we can make a heatmap to know which parts of the image are most important.
- Use different topological features like Euclidean Characteristics (EC) or persistence images.
- Obtain more data, especially for Her2, to increase the accuracies of classification.
- Analyze other components in the images like identifying cell types in addition to the nucleic information.

References

1. Rączkowski, Ł., Możejko, M., Zambonelli, J. et al. ARA: accurate, reliable and active histopathological image classification framework with Bayesian deep learning. Sci Rep 9, 14347 (2019). https://doi.org/10.1038/s41598-019-50587-1
2. Richard J. Chen, Ming Y. Lu, Drew F.K. Williamson, Tiffany Y. Chen, Jana Lipkova, Zahra Noor, Muhammad Shaban, Maha Shady, Mane Williams, Bumjin Joo, Faisal Mahmood, Pan-cancer integrative histology-genomic analysis via multimodal deep learning, Cancer Cell, Volume 40, Issue 8, 2022, Pages 865-878.e6, ISSN 1535-6108, https://doi.org/10.1016/j.ccell.2022.07.004.
3. Singh, N., Couture, H.D., Marron, J.S., Perou, C., Niethammer, M. (2014). Topological Descriptors of Histology Images. In: Wu, G., Zhang, D., Zhou, L. (eds) Machine Learning in Medical Imaging. MLMI 2014. Lecture Notes in Computer Science, vol 8679. Springer, Cham. https://doi.org/10.1007/978-3-319-10581-9_29
4. W. Lu, S. Graham, M. Bilal, N. Rajpoot and F. Minhas, "Capturing Cellular Topology in Multi-Gigapixel Pathology Images," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, 2020, pp. 1049-1058, doi: 10.1109/CVPRW50498.2020.00138.
5. Genomic Data Commons, National Cancer Institute. https://portal.gdc.cancer.gov

Acknowledgments

I want to thank my mentor, Dr. Crawford, for his support and advice that helped me immensely with the project. I also want to thank the PLME Faculty for helping support my project through the SRA Award.