Unsupervised learning for the analysis of cancer data

Allison J. Dusek[1,2], Alicia Bilbao Martinez[1], Alexander Claman[1], Daniel Bilbao Cortes[3], Vanessa Aguiar-Pulido[1]

[1]Department of Computer Science, University of Miami, Coral Gables, FL, USA

[2]Department of Mathematics and Computer Science, Samford University, Birmingham, AL, USA

[3]Department of Pathology and Laboratory Medicine, Sylvester Comprehensive Cancer Center, University of Miami Miller School of Medicine, Miami, FL, USA

Introduction

In 2021 breast cancer became the most common cancer globally with more than 3.8 million women reporting a history of breast cancer. About 43,250 women in the U.S. are expected to die in 2022 from breast cancer, yet the overall deaths have decreased by 1% between 2013 and 2018 as a result of advancements in treatment and early diagnosis though screening. [1] The advances in slide scanning technology and decrease in digital storage cost have made the process of attaining histopathologic stained tissue and its digitalization more feasible in recent years. The amount of available data is extensive and not all the underlying patterns present in the images are obvious by humans. Each slide contains in the order of thousands of cells and not all of them are cancerous, therefore identifying characteristics that reveal which ones are malignant or which tumors are recurrent, for example, are very relevant problems. However, before any analysis is performed, we must make sure that the data are homogeneous to ensure optimal results.

Machine learning techniques are used to collect and analyze data in an automated manner. In this work we will use unsupervised learning approaches to perform an exploratory analysis of the data. The goal of unsupervised learning is to discover hidden and interesting patterns in unlabeled data. We will also investigate the use of dimensionality reduction techniques to reduce the search space.

Methods

Data

The Breast Cancer Histopathological Image Classification (BreakHis) dataset[2] is composed of 4,960 microscopic images of breast tumor tissue collected from 82 patients using different magnifications. The dataset contains 588 benign and 588 malignant samples for the 40X magnification.

Techniques

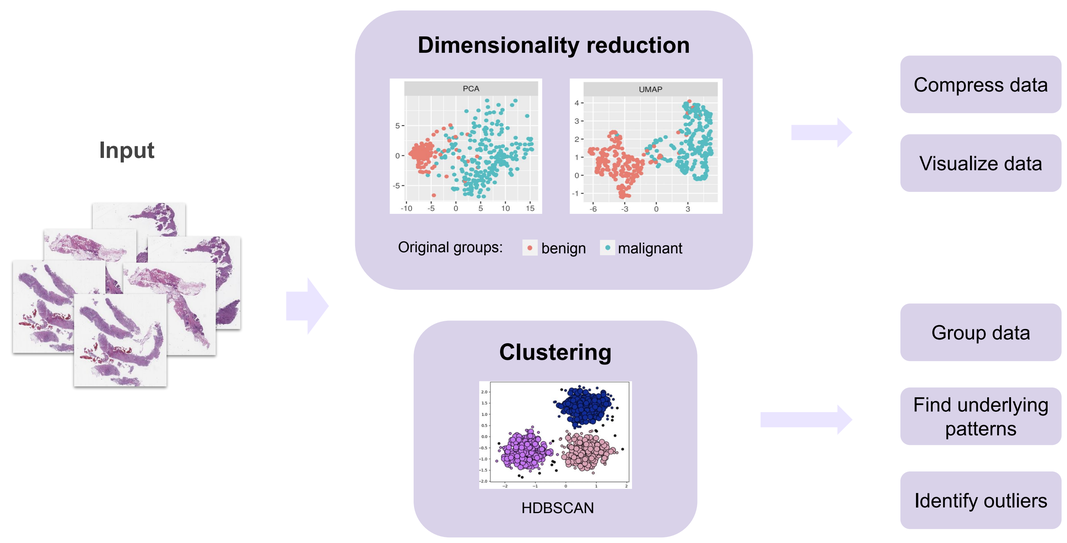
Dimensionality reduction:

- Goal: reduce the number of features in a dataset without losing information. Commonly used to summarize the data and to be able to plot it on a visual graph.
- In this work, PCA[3] (linear) and UMAP[4] (non-linear) were used to compress and visualize the data.

Clustering:

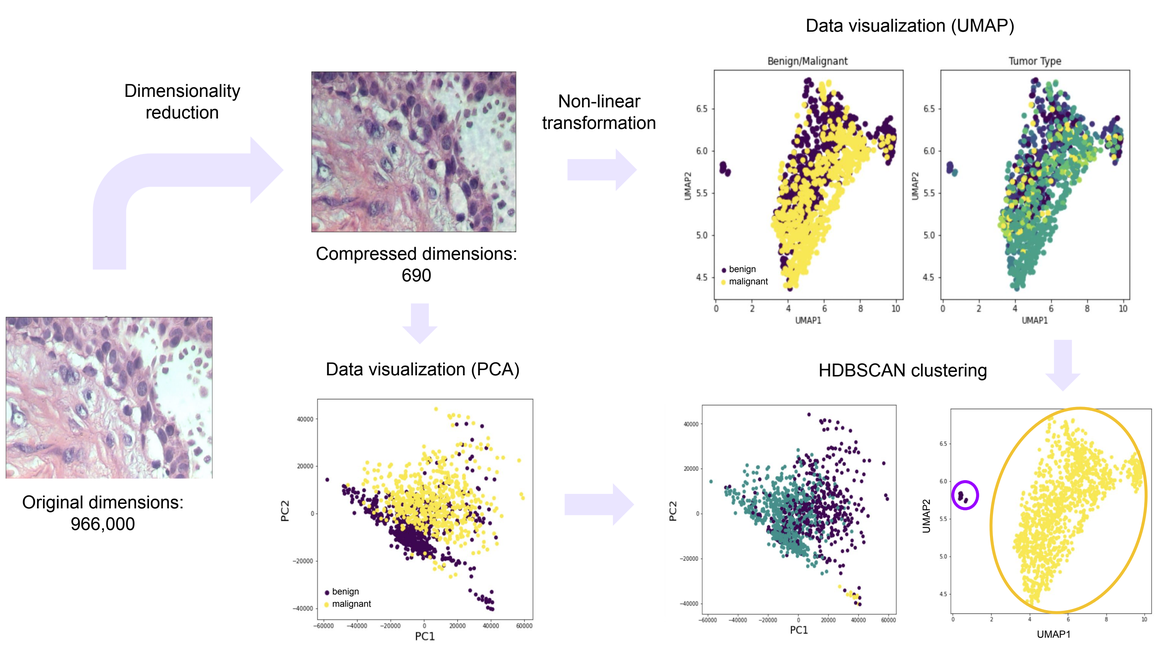
- Goal: learn patterns from unlabeled data. The data are separated into groups called clusters.
- In this work, HDBSCAN[5] was employed.

Figure 1. Exploratory analysis overview



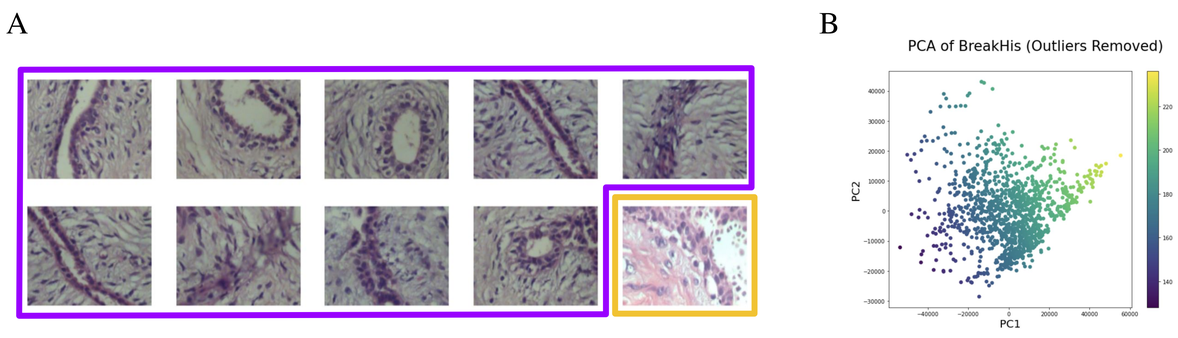
Results

Figure 2. Dimensionality reduction, visualization and clustering of breast cancer data



The exploratory analysis performed shows that there is a group of images that represent outliers. If we compare these images to those outside of the group, there is a humanly recognizable difference in backlighting.

Figure 3. Impact of intensity variation on the dataset. (a) Outliers (purple) vs. other images in the dataset (yellow). (b) PCA without outliers plotted on an intensity scale.



Even when the outliers are removed, we can clearly see that the difference in intensity is driving the first component of PCA.

Conclusions

- We applied two different dimensionality reduction techniques, as well as a clustering algorithm and visualized the data in a two-dimensional space.
- We were able to reduce the required computational resources by compressing the input data with PCA.
- We identified a group of outliers, therefore highlighting the need for exploratory analyses before building classification models.

Future work

- Consider different normalization techniques to avoid the need of removing outliers.
- Evaluate various image processing techniques to enhance the input data.
- Assess the impact of utilizing different scanners on the images generated.
- Determine whether other available data can be incorporated into the dataset that will be used as input for a classification algorithm.

Acknowledgements

Support for the program REU Site: Scientific Computing for Structure in Big or Complex Datasets is by the NSF grant CNS-1949972, program solicitation NSF 13-582, Research Experiences for Undergraduates (REU).

References

1. Breast Cancer Facts and Statistics. (2022). Breast Cancer. https://www.breastcancer.org/facts-statistics. Last accessed: 07/22/22.
2. Spanhol F.A. et al. (2015) A dataset for breast cancer histopathological image classification. IEEE Trans. Biomed. Eng., 63, 1455–1462.
3. Jollife, I.T. (2002). Principal Component Analysis, Second edition, New York: Springer-Verlag New York, Inc.
4. McInnes et al. (2018). UMAP: Uniform Manifold Approximation and Projection. J Open Source Softw, 3(29), 861.
5. McInnes L, Healy J, Astels S (2017). Hdbscan: hierarchical density based clustering. J Open Source Softw 2(11).