CSP 571 – Project

11/30/2024

Classification of Lung Cancer Images

Henrique Magalhaes Rio – A20544694

Riyaz Ahmed Mohammed – A20547233

1. The Data

For our project we choose the dataset that contains 15000 histopathological images of lung cancer, and it is divided in 3 classes each with 5,000 images. The 3 classes shown below are benign tissue which means no cancer, squamous cell carcinoma and adenocarcinoma. On *figure 1* we can see that there is a stark difference between the colors, and structure of the cells. The squamous nuclei and tissue seen to be a lot more packed together while the others are more spread out, the adenocarcinoma and benign tissue also show quite a bit of difference in the neclei structure. In this project we will attempt to use machine learning classification models along with dimension reduction models to try and classify different types of cancer.

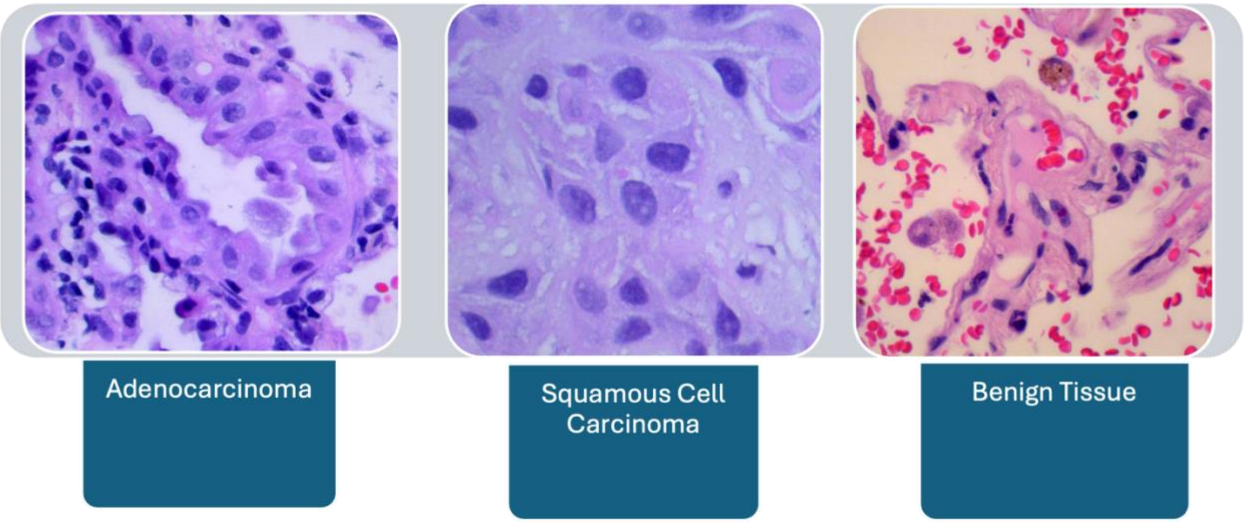

Figure 1 – Cancer types

2. Data Preparation

First, to prepare we start with the image $I_{768x768}$, then as we load each image and resize them to get a reduce image R_{64x64} for computational reasons. Next, we **grayscale** the images to remove the RGB channel. Finally, we use the function *np.flatten()* to get an array x_{4096} from the image R_{64x64}, which we will add for dataset X. In the end we have a dataset $X_{15000x4096}$ where each row represents an image, and each column represents a pixel position. For training and validation purposes we split the dataset X into an 80% training, 20% validation and 20% testing.

3. Dimension Reduction

To make the dataset more manageable in terms of computational time, we decided to do a principal component analysis (PCA), and use the first 2 principal components as the features in the data. We first started with the PCA for the RBG images which gave us a clean separation between the cancerous and benign tissue images as we can see in *figure 2*.

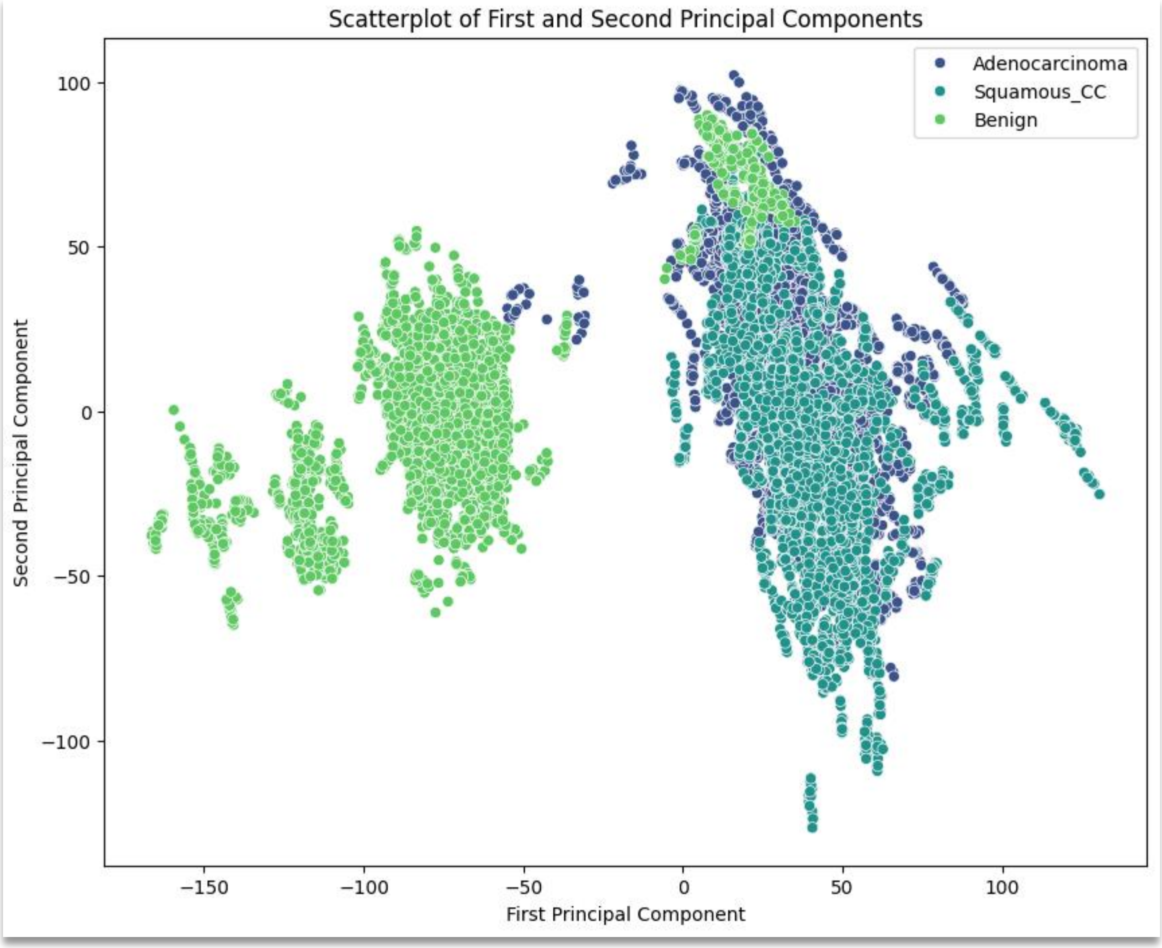

Figure 2 - RGB PCA

However, upon inspection of the individual principal components we realized that this separation was artificial as the principal component as using color as the most indicative factor not the structure of the cell. In *figure 3* we use the inverse transform to project the principal components into a 64x64 picture, in the rgb images its immediately clear that the biggest difference is on the colors, while in the grayscale as we increase the amount of principal components projected we see that the structure is being differentiated. Therefore, for the predictions we used the grayscale dataset since we are interested in the structure of the tissue.

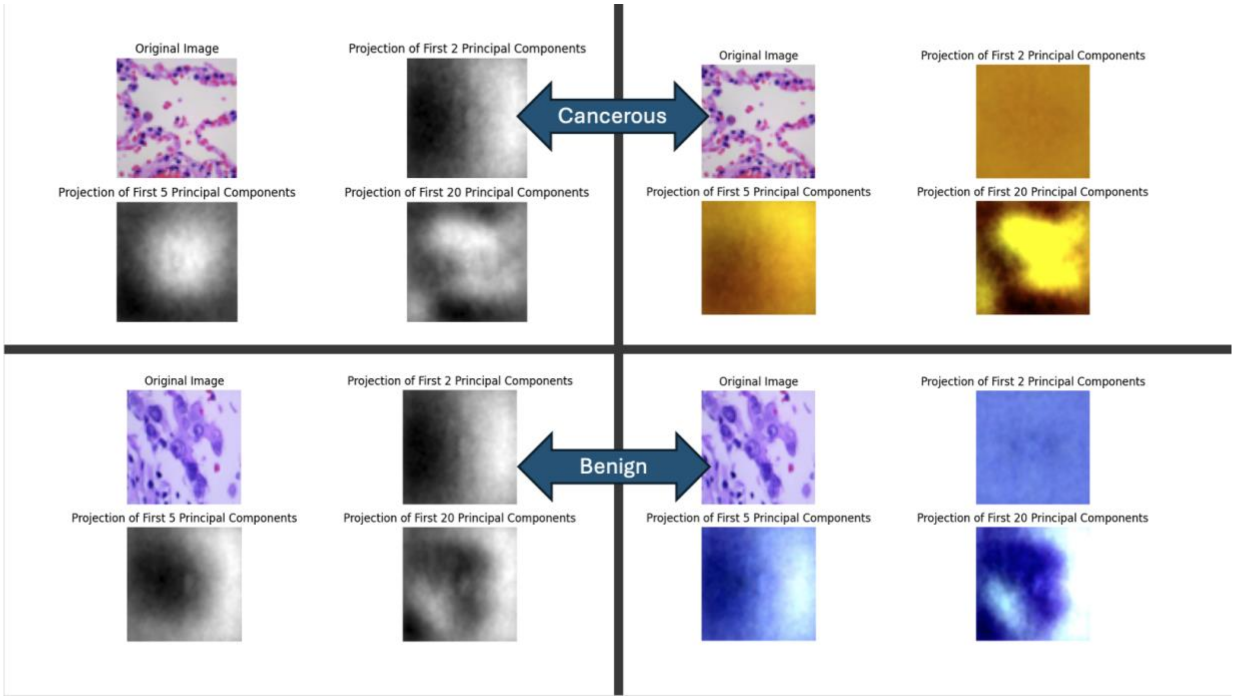
Original Image
Projection of First 2 Principal Components
Cancerous
Original Image
Projection of First 2 Principal Components
Projection of First 5 Principal Components
Projection of First 20 Principal Components
Projection of First 5 Principal Components
Projection of First 20 Principal Components
Original Image
Projection of First 2 Principal Components
Benign
Original Image
Projection of First 2 Principal Components
Projection of First 5 Principal Components
Projection of First 20 Principal Components
Projection of First 5 Principal Components
Projection of First 20 Principal Components

Figure 3 - PCA Projections

In *figure 4* we can visualize the first and second principal components of the grayscale images, while not as divided as the RGB images there is still some separations between the benign and cancerous tissues. However, things get more mixed when trying to discern between the 2 types of cancerous tissue.

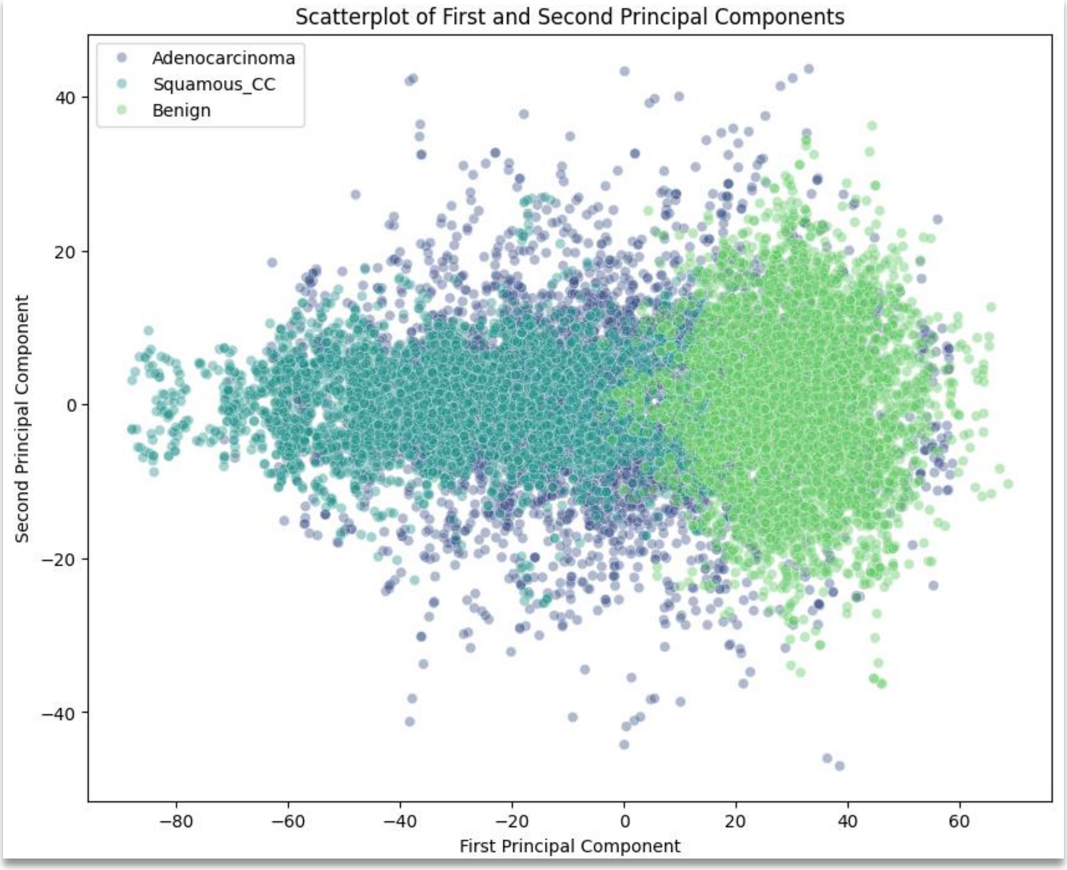

Figure 4 – PCA of grayscale images.

4. Classification Models

4.1 Decision Tree

The first model we used for

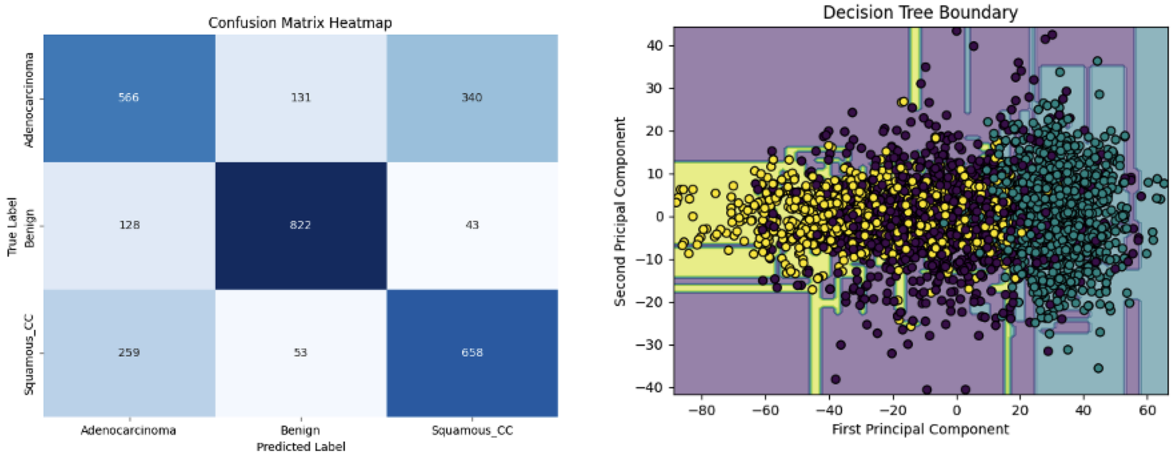

Figure 5 - DT Boundary and CM

Code Repository

https://github.com/henriquem27/csp571_Project

Work Cited

Borkowski AA, Bui MM, Thomas LB, Wilson CP, DeLand LA, Mastorides SM. Lung and Colon Cancer Histopathological Image Dataset (LC25000). arXiv:1912.12142v1 [eess.IV], 2019