CSP 571 – Project – Group 48

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 nuclei 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.

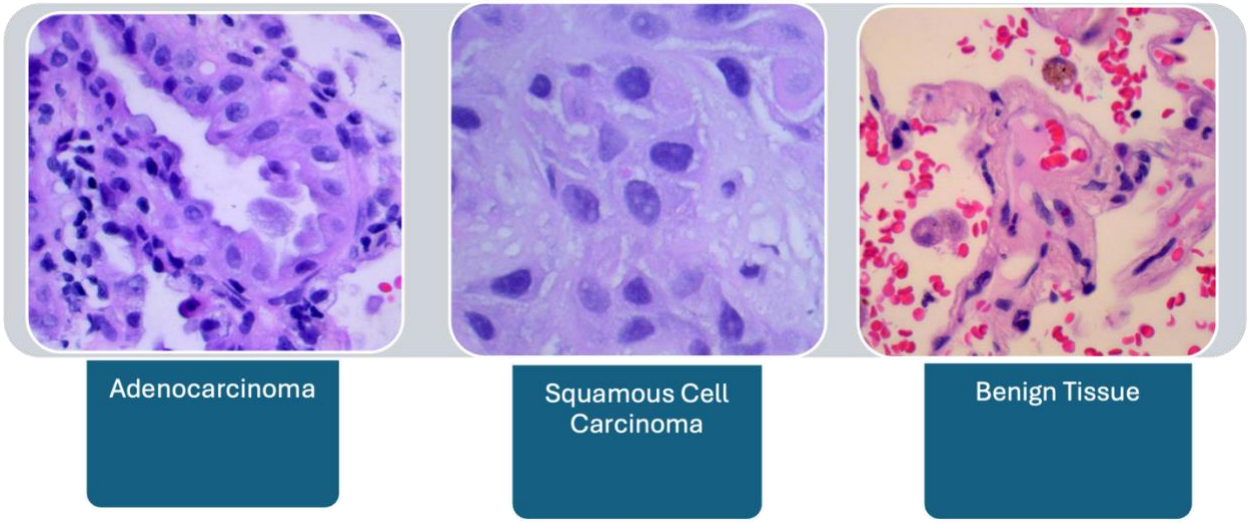

Figure 1 – Cancer types

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*.

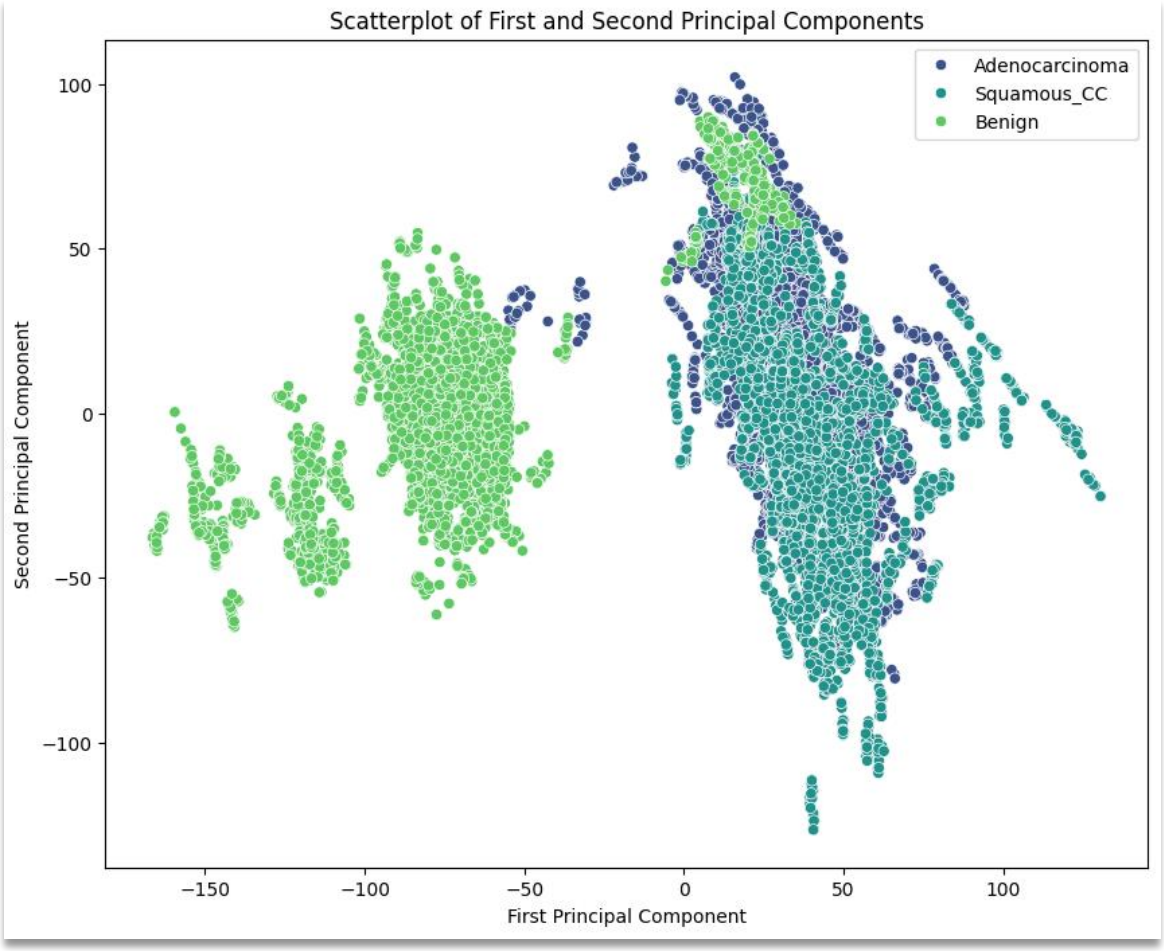

Figure 2 - RGB PCA

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.

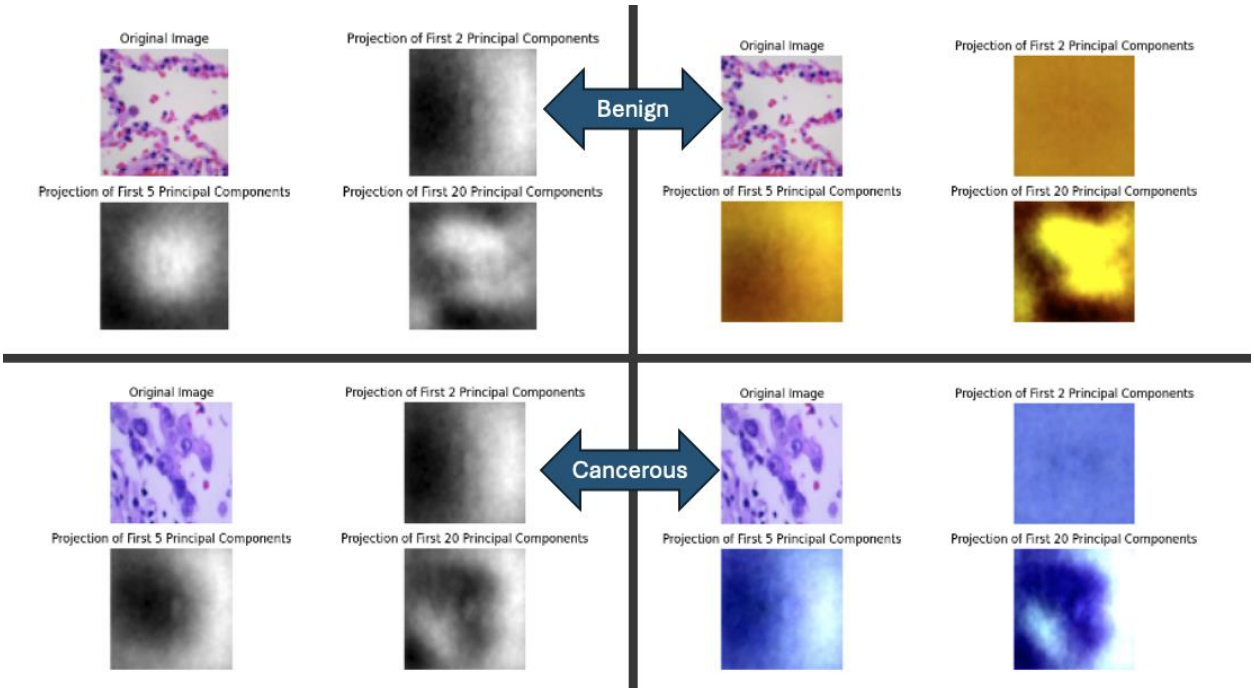

Figure 3 - PCA Projections

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.

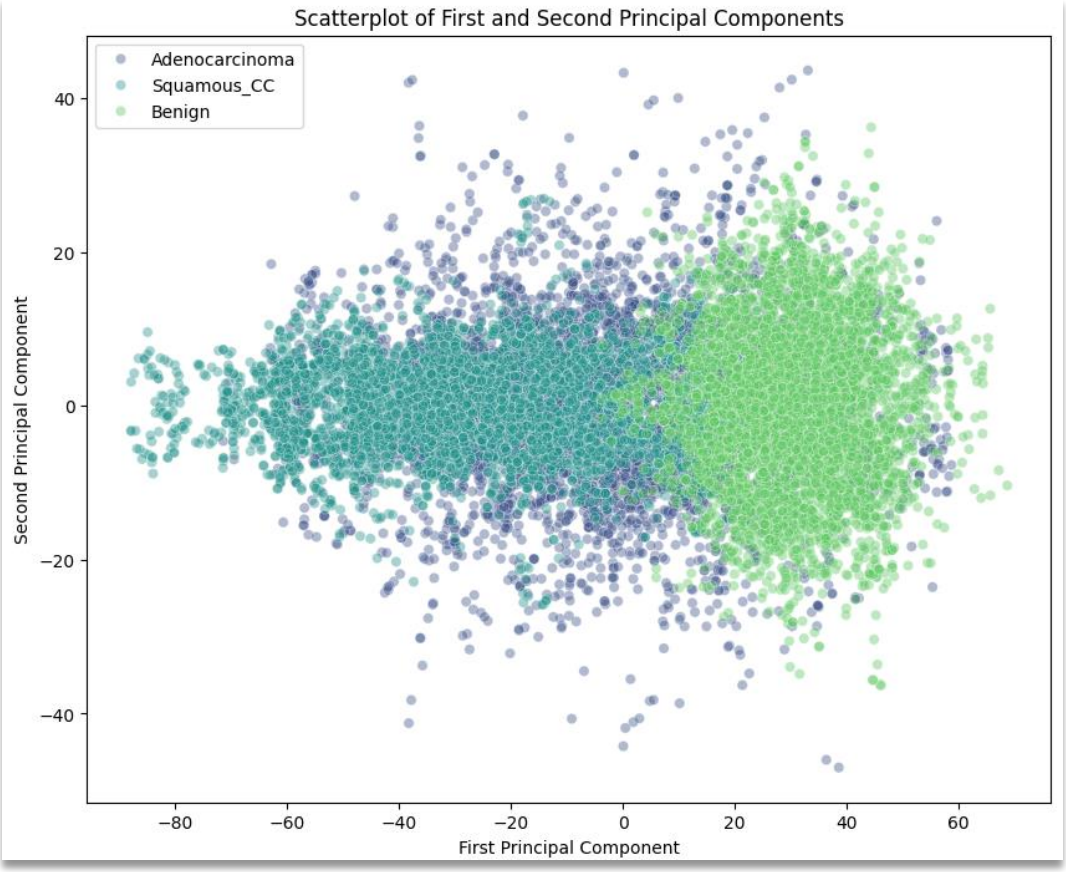

Figure 4 – PCA of grayscale images.

4. Classification Models

With the 2 PCS we tried to train and fit different kinds of models, the main score we are looking is the overall accuracy. However, in this case the precision ($Precision = \frac{TP}{TP+FP}$) and the false positive rate (FPR) are also of importance, because classifying a patient with a benign tissue when they have lung cancer could potentially be dangerous.

4.1 Decision Tree

The first model we used for classification was the decision tree using the first 2 principal components. In *figure 5,* we can see that the boundaries are really squared out, which leads to poor accuracy of around 0.68 on the test dataset. However, despite the low

accuracy the decision tree model still had a decent precision (0.82) in predicting benigh tissue as we can see from *table 1*.

Table 1 - DT Report

	precision	recall	f1-score	support
Adenocarcinoma	0.59	0.55	0.57	1037
Benign	0.82	0.83	0.82	993
Squamous_CC	0.63	0.68	0.65	970
accuracy			0.68	3000
macro avg	0.68	0.68	0.68	3000
weighted avg	0.68	0.68	0.68	3000

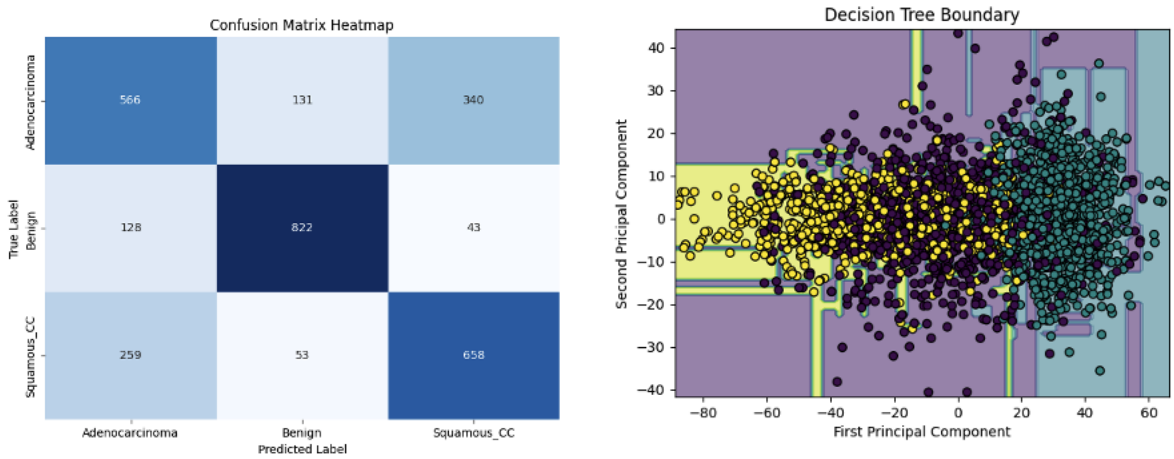

Figure 5 - DT Boundary and CM

4.2 KNN

We trained a KNN model with the number of neighbors from 1 to 30 and got that the model with 9 neighbors had the best validation accuracy with 0.718333. In *figure 6*, we can see that the decision boundary is more round which makes a little better at capturing the data. In *table 2,* we can see we have better overall accuracy when compared to the decision tree model, but worse precision of benign tissue.

Table 2 - KNN Report

	precision	recall	f1-score	support
Adenocarcinoma	0.62	0.52	0.57	1037
Benign	0.81	0.90	0.85	993
Squamous_CC	0.65	0.68	0.66	970
accuracy			0.70	3000
macro avg	0.69	0.70	0.69	3000
weighted avg	0.69	0.70	0.69	3000

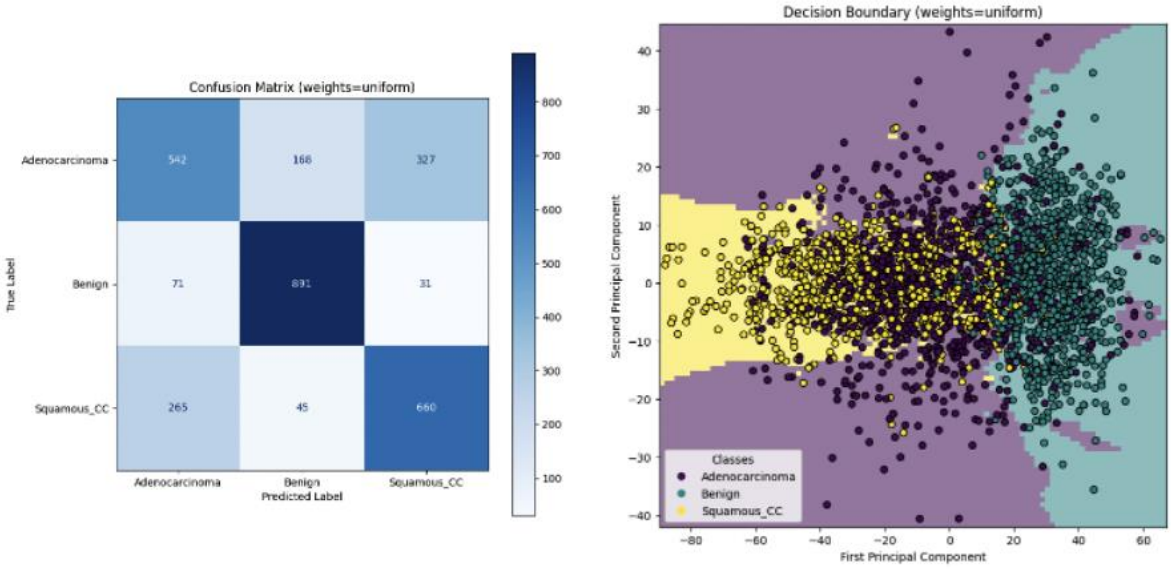

Figure 6 - KNN CM and DB

4.3 Support Vector Machine (SVM)

For the SVM, we decided to use a polynomial kernel to try to get a more curved boundary as we can see on *figure 7*. As we can see from the table the accuracy is lower when compared to KNN model, however, we have a higher precision on benign tissue with 0.86. Also, looking at the confusion matrix heatmap we manage to significantly reduce the benign false positives.

Table 3 – SVM report

	precision	recall	f1-score	support
Adenocarcinoma	0.49	0.81	0.61	1037
Benign	0.86	0.74	0.79	993
Squamous_CC	0.81	0.35	0.49	970
accuracy			0.64	3000
macro avg	0.72	0.64	0.63	3000
weighted avg	0.72	0.64	0.63	3000

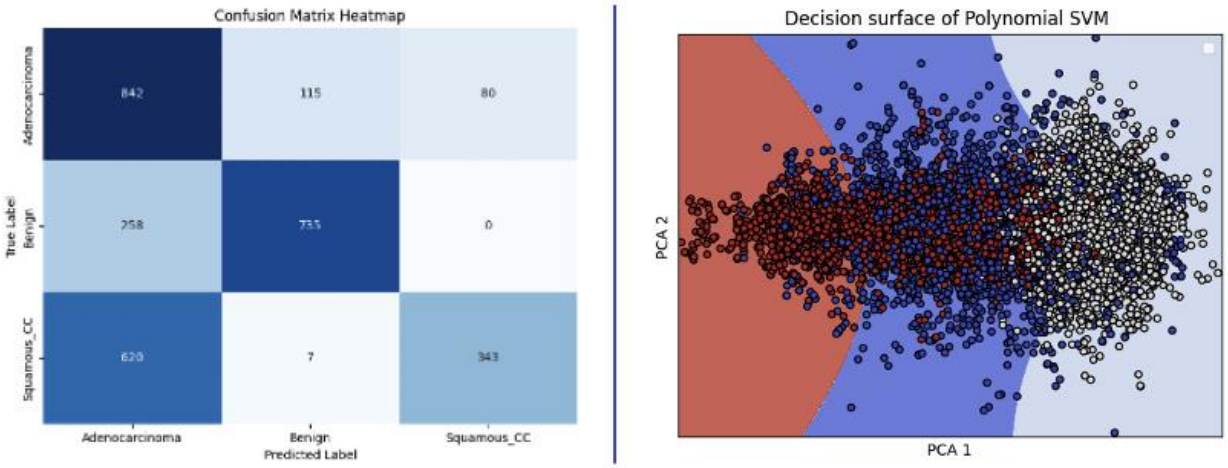

Figure 7 – SVM CM and DB

4.4 Random Forest (2 Principal Components)

We first trained a model on the training and validation set using the parameter grid on *figure 8*. Based on the results, we found that the best models had a max depth of 30, number of estimators of 500, and minimum samples split of 20 with validation accuracy of 0.733.

```
param_grid = {
    'n_estimators': [ 250,500, 1000],
    'max_depth': [10,20, 30, 50],
    'min_samples_split': [10,20, 30, 40]
}
```

Figure 8 – RF forest

Looking at the results in *table 4* we can see a minimal improvement in terms of overall accuracy, when compared to the other models. However, when in terms of the precision of benign tissue seems have dropped.

Table 4 - RF Report

	precision	recall	f1-score	support
Adenocarcinoma	0.64	0.54	0.58	1037
Benign	0.81	0.91	0.86	993
Squamous_CC	0.66	0.69	0.67	970
accuracy			0.71	3000
macro avg	0.70	0.71	0.71	3000
weighted avg	0.70	0.71	0.70	3000

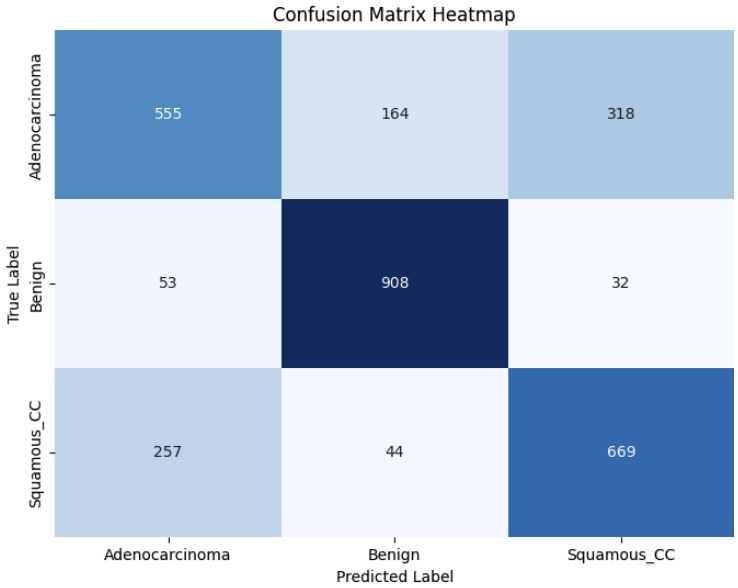

Figure 9 – RF CM

4.5 Random Forest (50 Principal Components)

Since random forest does better with higher dimensionality, we tried increasing the number of principal components to 50 to compare the results. And as we can see from *table 5*, we have a significant improvement in the overall accuracy when compared to the other models. However, it still has not reached the level of precision of SVM.

Table 5 – Report for RF

	precision	recall	f1-score	support
Adenocarcinoma	0.82	0.76	0.79	1037
Benign	0.84	0.94	0.89	993
Squamous_CC	0.90	0.88	0.89	970
accuracy			0.85	3000
macro avg	0.86	0.86	0.85	3000
weighted avg	0.86	0.85	0.85	3000

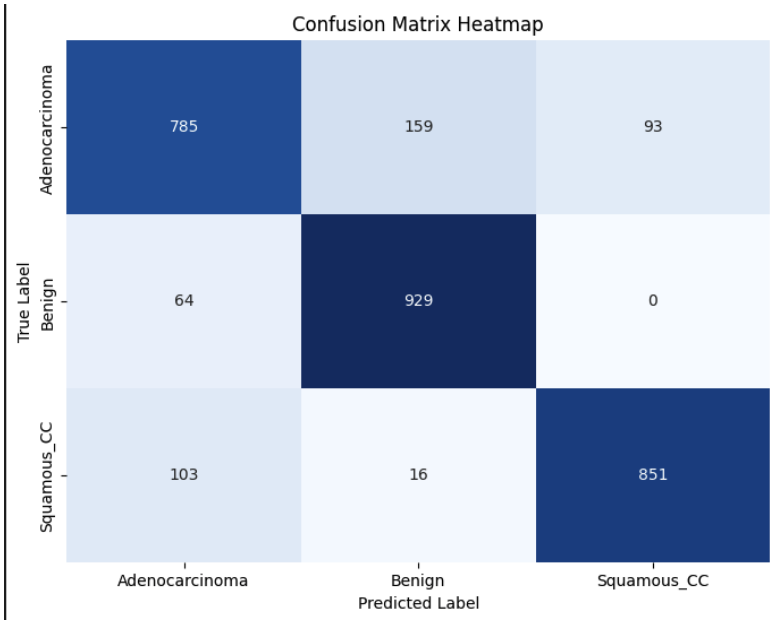

Figure 10 – RF CM 50PC

5. Conclusion

Table 6 - Summary of Findings

Model	Overall Accuracy	Precision of Benign
Decision Tree (2 PC)	0.68	0.82
Decision Tree (50 PC)	0.78	0.83
KNN (2 PC)	0.70	0.81
KNN (50 PC)	0.70	0.81
SVM (2 PC)	0.64	0.86
SVM (50 PC)	0.80	0.89
Random Forest (2 PC)	0.71	0.81
Random Forest (50 PC)	0.85	0.84

Overall, given the limitations imposed on the data by using the first 2 principal components of such large dataset the results are interesting particularly the fact that in terms of precision the polynomial SVM achieved the best overall result. We also increased the number of principal components to 50 for all models for the sake of comparison as presented by the results in *table 6*. As we can see not all models benefit from the increase in the dimensions of the data, KNN for instance had no improvements in the accuracy and precision of benign tissue. However, SVM had a massive improvement in in overall accuracy and decent improvement in precision making it a much closer contender to the random forest model.

Code Repository

https://github.com/henriquem27/csp571_Project

Dataset

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