COMP 448/548 – Medical Image Analysis Homework #2

Part I)

Before clustering the cells, the first step was to extract the features of each patch in the image. It was important to find a reasonable patch size at first; we manually tried and found a number that fit cells satisfactorily. For each patch, after intensity-based features are calculated, we also added a gray-level cooccurrence matrix in order to preserve spatial relations between pixel values. Finally, from the matrix, we calculate angular second moment, maximum probability, inverse difference moment, and entropy to represent the textural features in the clustering algorithm.

Part II)

Feature normalization is crucial in the context of the k-means clustering algorithm due to several reasons. First, normalizing features ensures scale invariance, where features with different scales contribute proportionately to the clustering process. This prevents biased cluster assignments based on the original scale of the features. Second, normalization promotes convergence and stability of the algorithm by preventing features with larger scales from dominating the clustering process. It ensures that all dimensions are equally treated and prevents any particular feature from having undue influence.

In addition to feature normalization, addressing the class imbalance problem is vital in data analysis. The task at hand involves determining the total number of inflammatory, epithelial, and spindle-shaped cells in all images. Class imbalance can lead to biased clustering, with the algorithm favoring the majority class and overlooking important patterns in minority classes. To tackle this issue, a weighted clustering approach has been implemented using the weighted clustering function.

The weighted_clustering() function takes two arguments: all_features, representing the feature vectors, and cell_dict, a dictionary containing the counts of cells for each class. The function calculates weights for each class based on the inverse of the class count. This ensures that classes with fewer samples have higher weights, while classes with more samples have lower weights. The weights are then normalized, to sum up to 1. The feature vectors in all features are multiplied element-wise with their corresponding weights, resulting in a weighted dataset.

By applying this weighted clustering approach, the algorithm can mitigate the class imbalance problem by giving more importance to the minority classes during clustering. This allows for a

more balanced representation of classes and improves the accuracy of clustering, particularly when dealing with imbalanced datasets.

The incorporation of normalization and weighted class clustering techniques has led to an improvement in accuracy from 0.75 to 0.8. This increase indicates that these techniques have positively impacted the clustering performance, resulting in more accurate and meaningful clustering results. It's important to consider other evaluation metrics alongside accuracy to comprehensively assess clustering performance. Nevertheless, this improvement demonstrates the effectiveness of the applied techniques in addressing feature scaling and class imbalance challenges for the given task.

To classify each cell, we extracted features from the cropped image, then trained the k-means model based on that information. As it was mentioned in the instructions as a class imbalance problem, the results were not satisfying. To properly represent the minority data, we used weighted clustering to scale the feature vectors based on their quantities. This extra method achieved a greater accuracy in clustering. (0.68 to 0.72 in our metric)

Part III)

Three parameters were investigated: N, binNumber, and d. Through experimentation and evaluation, several observations were made. Firstly, the N value, which represents the size of the neighborhood considered for feature extraction, was found to have a significant influence on clustering accuracy. Notably, N values of 16 and 32 consistently yielded the highest accuracy results, indicating the importance of capturing the appropriate neighborhood context for feature extraction.

Secondly, the co-occurrence matrix, a commonly used feature in image analysis tasks, was found to have a limited effect on clustering accuracy. This suggests that other features or metrics may play a more crucial role in determining accurate clustering results.

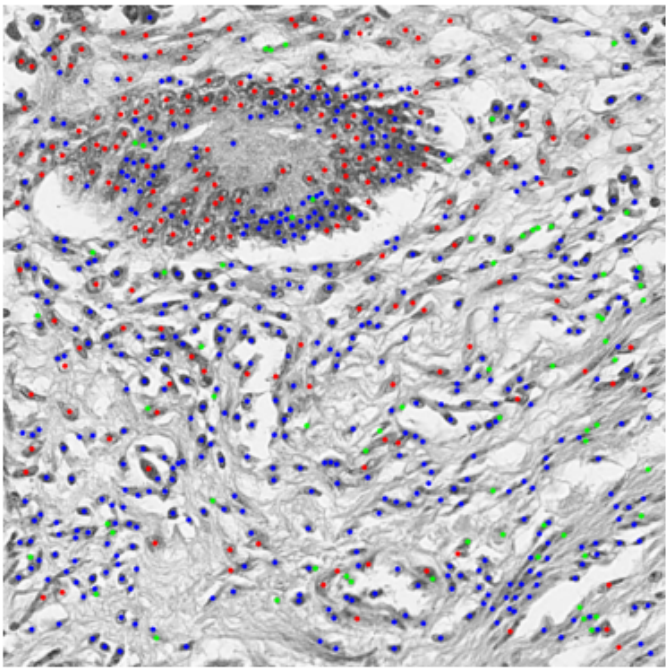
Lastly, the binNumber and d values were observed to have a smaller impact on clustering accuracy compared to the N value. After thorough experimentation, binNumber values of 10 and 30 were selected as they provided satisfactory accuracy while considering computational efficiency.

Overall, the results indicate that the N value is the most critical parameter for achieving optimal clustering accuracy. The findings highlight the importance of considering the neighborhood size when performing feature extraction. Additionally, the limited impact of the co-occurrence matrix suggests the need to explore other features or metrics for improving clustering accuracy. The chosen binNumber values strike a balance between accuracy and computational efficiency. These conclusions provide valuable insights for parameter selection in feature extraction and its impact on clustering accuracy.

Experiment Results:

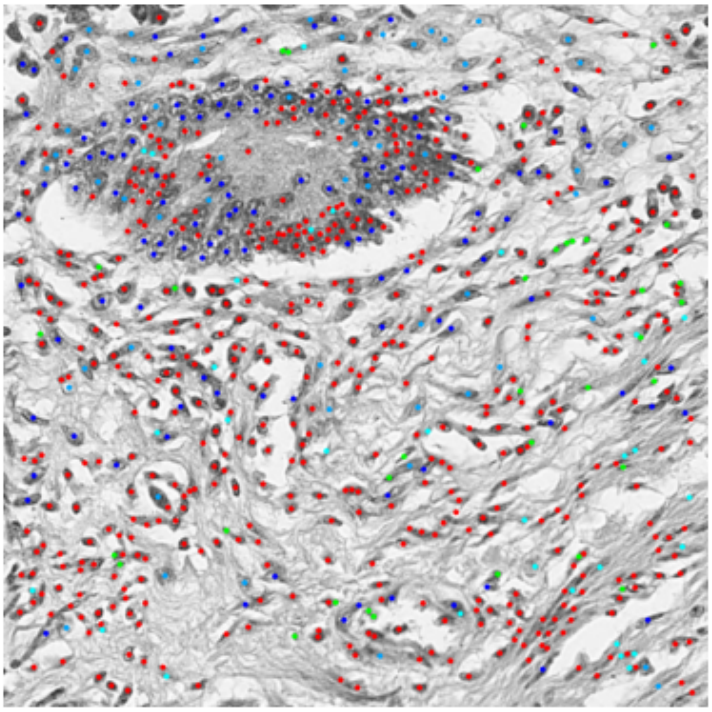
Parameters	Bin Number	d	N	k
	10	1	18	3

Cluster	Inflammation	Epithelial	Spindle
0	0.12	0.21	0.66
1	0.00	0.08	0.92
2	0.01	0.44	0.56



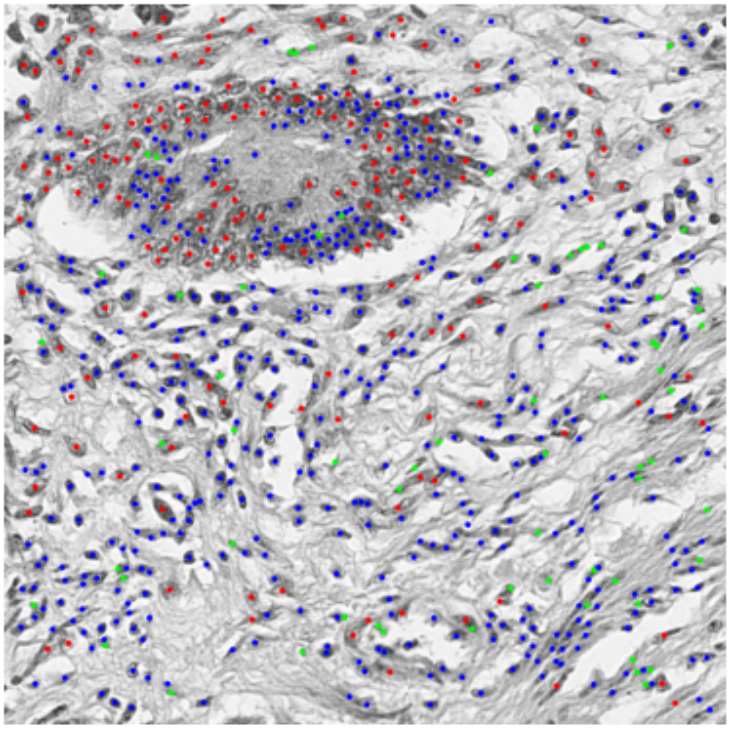
Parameters	Bin Number	d	N	k
	10	1	18	5

Cluster	Inflammation	Epithelial	Spindle
0	0.01	0.52	0.47
1	0.00	0.03	0.97
2	0.12	0.22	0.66
3	0.00	0.13	0.87
4	0.00	0.23	0.77



Parameters	Bin Number	d	N	k
	10	1	36	3

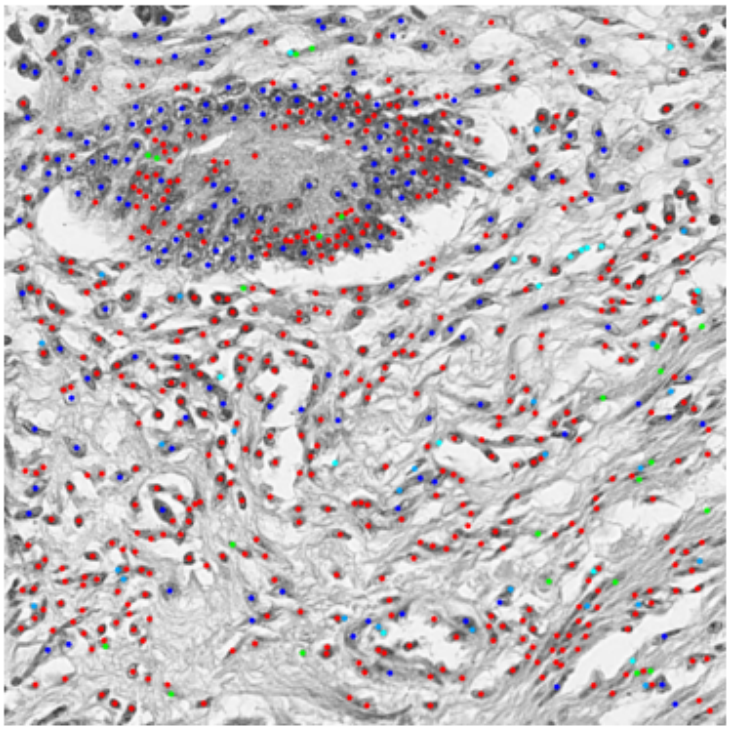
Cluster	Inflammation	Epithelial	Spindle
0	0.12	0.21	0.67
1	0.00	0.08	0.92
2	0.01	0.45	0.54



Parameters	Bin Number	d	N	k
	10	1	36	5

Cluster	Inflammation	Epithelial	Spindle
0	0.01	0.45	0.54
1	0.00	0.17	0.83

2	0.12	0.21	0.67
3	0.00	0.00	1.00
4	0.00	0.04	0.96



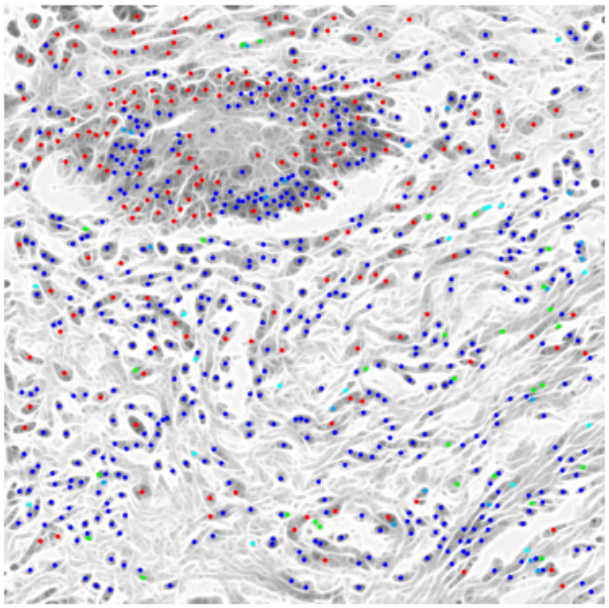
Part IV - Extension 1)

Gabor Filter

Parameters	Bin Number	d	N	k
	10	1	18	5

Cluster	Inflammation	Epithelial	Spindle

0	0.12	0.21	0.67
1	0.00	0.00	1.00
2	0.01	0.44	0.55
3	0.00	0.00	1.00
4	0.00	0.25	0.75



Parameters	Bin Number	d	N	k
	10	1	36	5

Cluster	Inflammation	Epithelial	Spindle
0	0.12	0.21	0.67
1	0.00	0.03	0.97
2	0.01	0.45	0.54
3	0.00	0.00	1.00
4	0.00	0.22	0.78