Comp448 Medical Image Analysis Homework 2 Report

Name – Surname: Barış Kaplan (**KU ID Number: 0069054)**

Experiments

For k = 3:

binNumber = 6, d = 5, N = 10:

Ratios of Inflammatory, Epithelial, and Spindle-Shaped Cells for Each Cluster (for train images):

	Inflammatory	Epithelial	Spindle-shaped
Cluster 1	0.05977382875605816	0.5137318255250404	0.42649434571890144
Cluster 2	0.6216216216216216	0.10135135135135136	0.27702702702702703
Cluster 3	0.00	0.37117903930131	0.62882096069869

Ratios of Inflammatory, Epithelial, and Spindle-Shaped Cells for Each Cluster (for test images):

	Inflammatory	Epithelial	Spindle-shaped
Cluster 1	0.001451378809869376	0.4731494920174166	0.525399129172714
Cluster 2	0.8190184049079755	0.05828220858895705	0.12269938650306748
Cluster 3	0.1256637168141593	0.30796460176991153	0.5663716814159292

Visuals of The Clustering Results:

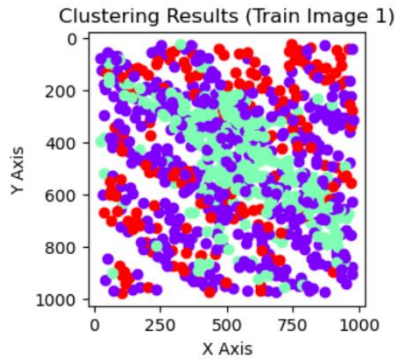

Figure 1: train_8.png Clustering Results

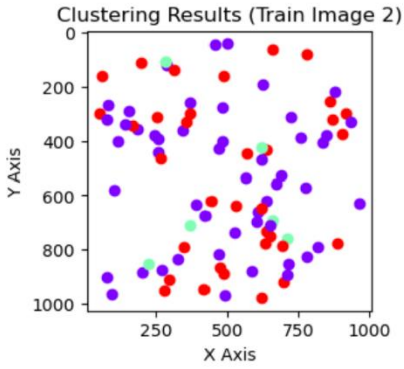

Figure 2: train_11.png Clustering Results

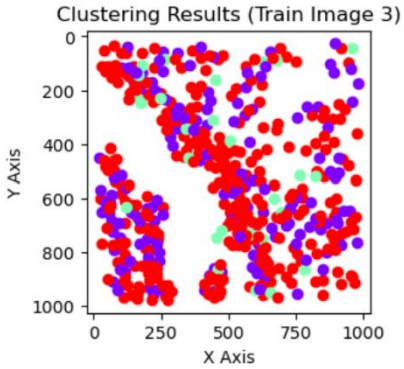

Figure 3: train_14.png Clustering Results

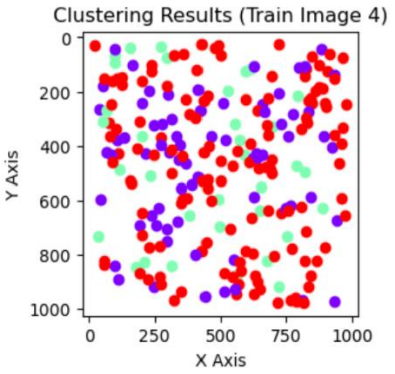

Figure 4: train_21.png Clustering Results

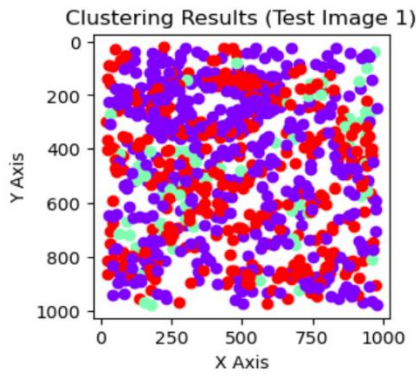

Figure 5: test_1.png Clustering Results

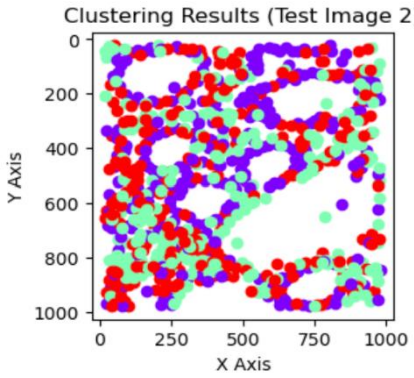

Figure 6: test_10.png Clustering Results

binNumber = 9, d = 5, N = 10:

Ratios of Inflammatory, Epithelial, and Spindle-Shaped Cells for Each Cluster (for train images):

	Inflammatory	Epithelial	Spindle-shaped
Cluster 1	0.059870550161812294	0.5129449838187702	0.42718446601941745
Cluster 2	0.00	0.37209302325581395	0.627906976744186
Cluster 3	0.6216216216216216	0.10135135135135136	0.27702702702702703

Ratios of Inflammatory, Epithelial, and Spindle-Shaped Cells for Each Cluster (for test images):

	Inflammatory	Epithelial	Spindle-shaped
Cluster 1	0.001455604075691412	0.47307132459970885	0.5254730713245997
Cluster 2	0.8170731707317073	0.057926829268292686	0.125
Cluster 3	0.12389380530973451	0.30973451327433627	0.5663716814159292

Visuals of The Clustering Results:

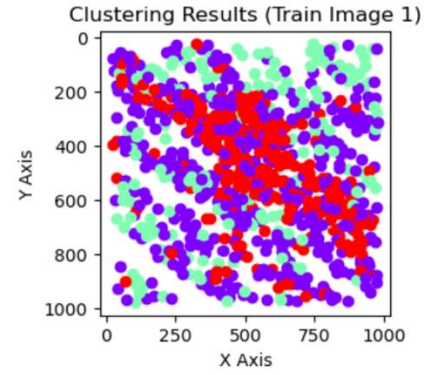

Figure 7: train_8.png Clustering Results

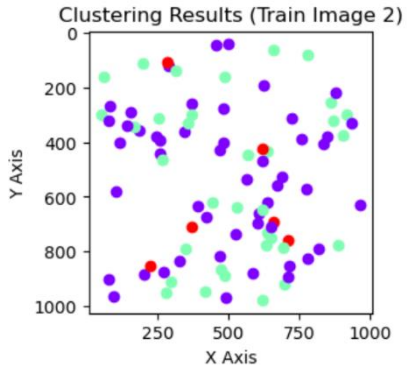

Figure 8: train_11.png Clustering Results

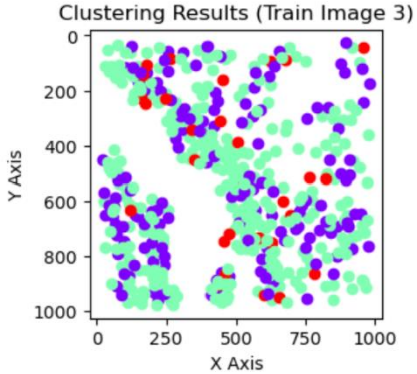

Figure 9: train_14.png Clustering Results

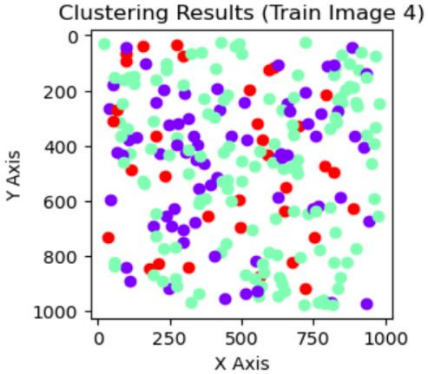

Figure 10: train_21.png Clustering Results

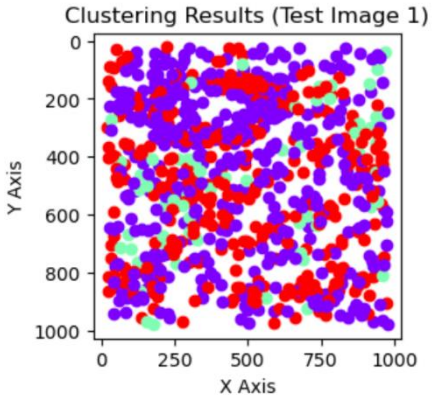

Figure 11: test_1.png Clustering Results

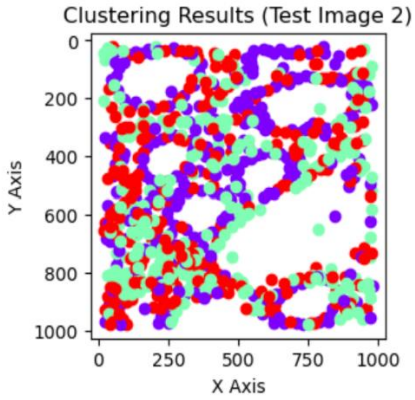

Figure 12: test_10.png Clustering Results

For k = 5:

binNumber = 6, d = 5, N = 10:

Ratios of Inflammatory, Epithelial, and Spindle-Shaped Cells for Each Cluster (for train images):

	Inflammatory	Epithelial	Spindle-shaped
Cluster 1	0.00	0.2631578947368421	0.7368421052631579
Cluster 2	0.07317073170731707	0.5392953929539296	0.3875338753387534
Cluster 3	0.004081632653061225	0.5204081632653061	0.47551020408163264
Cluster 4	0.7751937984496124	0.031007751937984496	0.1937984496124031
Cluster 5	0.39316239316239315	0.19230769230769232	0.41452991452991456

Ratios of Inflammatory, Epithelial, and Spindle-Shaped Cells for Each Cluster (for test images):

	Inflammatory	Epithelial	Spindle-shaped
Cluster 1	0.6280193236714976	0.10144927536231885	0.27053140096618356
Cluster 2	0.008146639511201629	0.4745417515274949	0.5173116089613035
Cluster 3	0.9065934065934066	0.027472527472527472	0.06593406593406594
Cluster 4	0.11627906976744186	0.29651162790697677	0.5872093023255814
Cluster 5	0.00	0.4438202247191011	0.5561797752808989

Visuals of The Clustering Results:

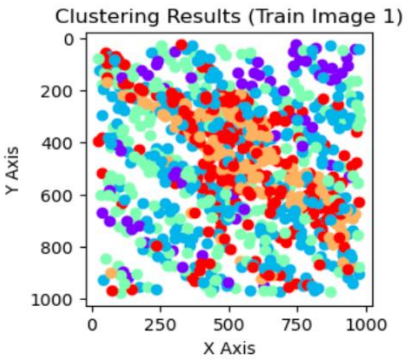

Figure 13: train_8.png Clustering Results

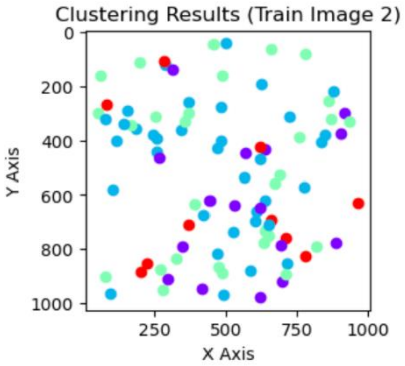

Figure 14: train_11.png Clustering Results

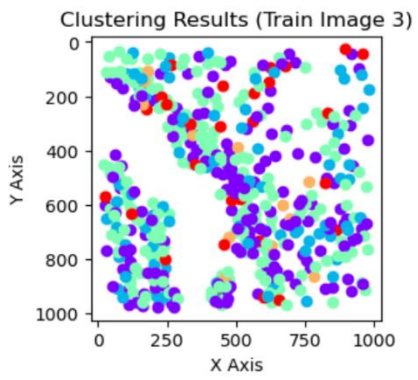

Figure 15: train_14.png Clustering Results

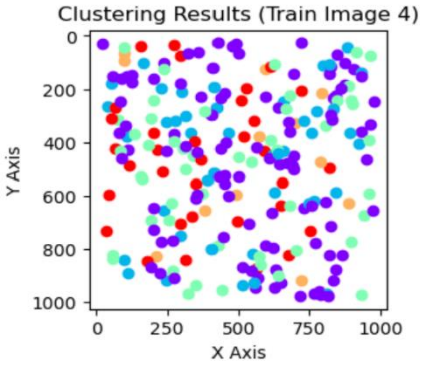

Figure 16: train_21.png Clustering Results

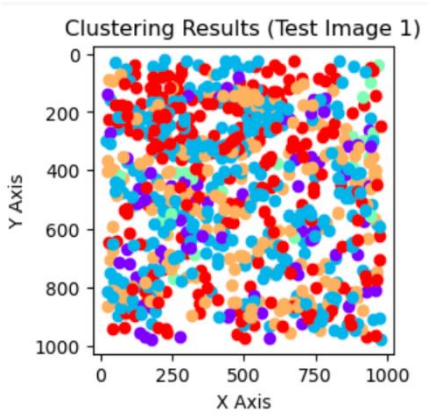

Figure 17: test_1.png Clustering Results

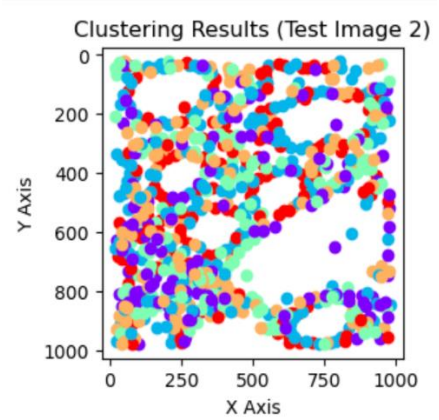

Figure 18: test_10.png Clustering Results

binNumber = 9, d = 5, N = 10:

Ratios of Inflammatory, Epithelial, and Spindle-Shaped Cells for Each Cluster (for train images):

	Inflammatory	Epithelial	Spindle-shaped
Cluster 1	0.00	0.516260162601626	0.483739837398374
Cluster 2	0.07105263157894737	0.5447368421052632	0.38421052631578945
Cluster 3	0.7769230769230769	0.03076923076923077	0.19230769230769232
Cluster 4	0.38589211618257263	0.1950207468879668	0.4190871369294606
Cluster 5	0.00	0.25348189415041783	0.7465181058495822

Ratios of Inflammatory, Epithelial, and Spindle-Shaped Cells for Each Cluster (for test images):

	Inflammatory	Epithelial	Spindle-shaped
Cluster 1	0.00	0.4425770308123249	0.5574229691876751
Cluster 2	0.9065934065934066	0.027472527472527472	0.06593406593406594
Cluster 3	0.12244897959183673	0.2915451895043732	0.5860058309037901
Cluster 4	0.6305418719211823	0.09852216748768473	0.270935960591133
Cluster 5	0.00808080808080808	0.4767676767676768	0.5151515151515151

Visuals of The Clustering Results:

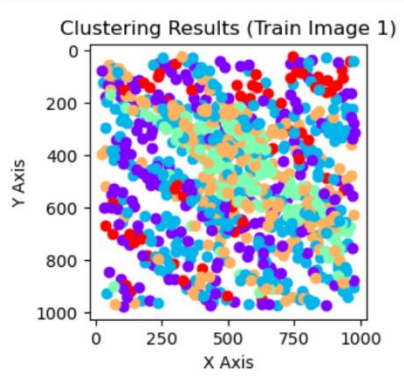

Figure 19: train_8.png Clustering Results

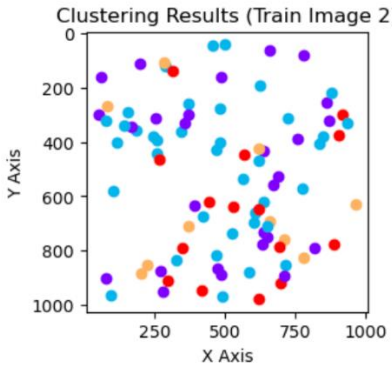

Figure 20: train_11.png Clustering Results

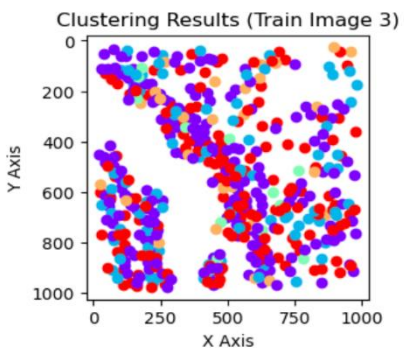

Figure 21: train_14.png Clustering Results

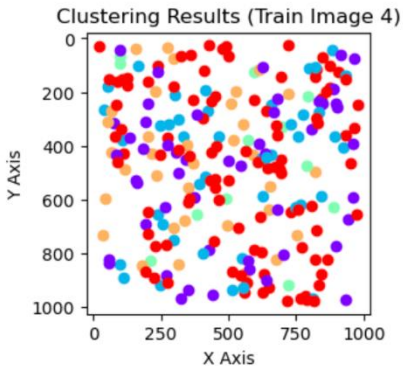

Figure 22: train_21.png Clustering Results

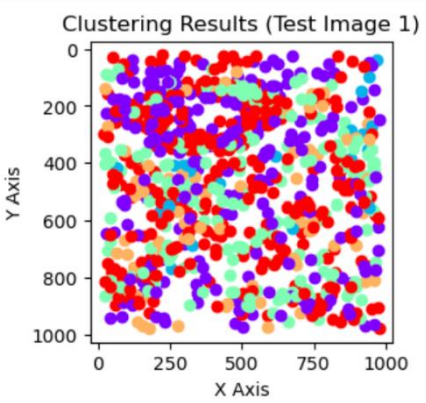

Figure 23: test_1.png Clustering Results

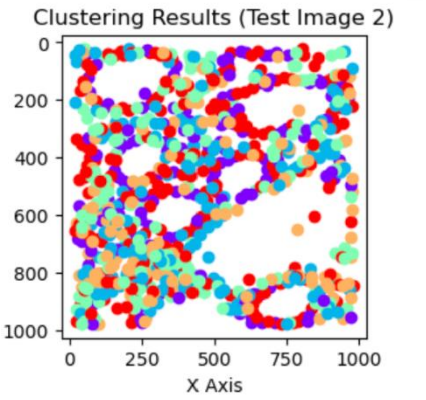

Figure 24: test_10.png Clustering Results

EXPLANATION AND INTERPRETATION OF THE EXPERIMENT RESULTS (including brief discussions about feature normalization, class imbalance, and how I selected the parameter values)

I have kept two scores for evaluating the performance of the k-means clustering: silhouette score and inertia score. Silhouette score measures the compactness of each cluster and the separation degree of the clusters. As this score approaches 1, the compactness of each cluster increases and the clusters become better separated from each other; which brings better clustering results. Therefore, I have selected the values of the binNumber, d, and N (window size) parameters such that they make the silhouette score as near to 1 as possible. The inertia score is another measure of clustering performance. In each cluster, It calculates the distance of a data point in that cluster to the centroid of that cluster. Then, it gets the square of each calculated distance. Finally, across a specific cluster, it sums up all the squared distances. In other words, it can be named as "in-cluster sum of squared distances". As the value of the inertia score increases, then the data points within a cluster become farther from each other, and therefore the in-cluster variability increases. Therefore, I tried to select the parameter values such that they make the inertia score as low as possible. Other than these, while selecting the parameter values, I paid attention to how compact each cluster is and how well the clusters are separated from each other by observing the visual results of the k-means clustering. In terms of these aspects, the parameter values that I used gave me better results. In the two combinations of parameters that I used, I only changed the value of the bin number in one combination and keep the other parameters the same. Because I want to better observe the specific effect of one parameter/factor on the results of k-means clustering. I have increased the value of the bin number from 6 to 9 in the second combination of parameters I used. The reasons for this were to improve the separation of clusters, obtain finer-grained clusters, obtain better cluster assignments, and enhance the sensitivity of the k-means clustering to the details and differences in data.

Overall, I obtained unequal distribution of cell types in each cluster and highly skewed clustering results for different parameter combinations, k values, and train/test images I used. For k = 3, I have used two combinations of the parameters which are binNumber, d, and N. The first combination of parameters I examined is binNumber = 6, d = 5, and N = 10. In addition, the second combination of parameters that I examined is binNumber = 9, d = 5, and N = 10. For the train images, when k = 3, binNumber = 6, d = 5, and N = 10; I have observed that the first cluster is dominated by the epithelial cells, then by the spindle cells. For this cluster, I have also observed that the inflammatory cells have the least proportion. For the second cluster, I observed that the cell ratios are ranked as inflammatory, spindle, and epithelial. For the third cluster, I observed that spindle-shaped cells had the greatest proportion, while inflammatory cells had the least at 0 percent. For the same cluster, I have also seen that the ratio of epithelial cells is in between the ratios of spindle-shaped and inflammatory cells. For the same parameters, when the images are the test images, I have observed that the first cluster is dominated by spindle-shaped cells, then by epithelial cells. For this cluster, I also concluded that the inflammatory cells have the least ratio. For the second cluster of test images, I concluded that while the inflammatory cells have the biggest proportion with domination, the epithelial cells have the least proportion. In comparison to the ratio of the other cell types, I also observed that the spindle-shaped cells have the middle ratio within this cluster. For the third cluster of the test images, I concluded that the spindle-shaped cells dominated. I also observed that the epithelial cells have the middle ratio in the third cluster, while the inflammatory cells take little place in this cluster.

I also utilized the parameter combination of binNumber = 9, d = 5, and N = 10. For these values, in the first cluster of the train images, I have seen that the ratios of the cells are ranked as epithelial, spindle, and inflammatory. In the second cluster of the train images, I concluded that the ratios of the cells are ranked as spindle, epithelial, and inflammatory. I also saw that the inflammatory cells take no place in this second cluster. For the third cluster of train images, I observed that the ratios of cells are ranked as inflammatory, spindle, and epithelial; where inflammatory cells dominated the cluster. For the test images, within the first cluster, I concluded that the ratios of the cells are ranked as spindle, epithelial, and inflammatory; where the inflammatory cells take almost no place. For the second cluster, I concluded that whereas the inflammatory cells have the biggest ratio, the epithelial cells have the least ratio. Furthermore, in the same cluster, I observed that the spindle-shaped cells have the middle amount of cell ratio. For the third cluster of the test images; I observed that the ratios of cells are ranked as spindle-shaped, epithelial, and inflammatory.

I have also conducted experiments for k = 5 with the same combinations of parameters. In this case; when the binNumber = 6, d = 5, and N = 10, and the images are train images, I observed that the cell ratios in the first cluster are ranked as spindle, epithelial, and inflammatory; where the inflammatory cells take no place. In the second cluster of the train images, I observed that the cell ratios are ranked as epithelial, spindle, and inflammatory. In the third cluster, I observed the same relationship of cell ratios as the second one but saw that the inflammatory cells take almost no place. For the fourth cluster, I observed that the ratios of cells are ranked as inflammatory, spindle, and epithelial; where inflammatory cells have a great amount of domination and epithelial cells take very little place. For the fifth cluster, I concluded that the spindle-shaped cells have the biggest cell ratio, while the epithelial cells have the least cell ratio.

For the same values of parameters, when the images are the test images, I saw that the ratios of cells are ranked as inflammatory, spindle, and epithelial. In addition, I observed that the biggest proportion of the second cluster is taken by spindle-shaped cells, and the smallest place in this cluster is taken by inflammatory cells. For Cluster 3, I observed that the inflammatory cells significantly dominated the cluster, while the other types of cells take almost no place. In the fourth cluster, I concluded that the spindle-shaped cells occupied the biggest proportion whereas the inflammatory cells took the least. For the fifth cluster, I observed that more than half of the cells are spindle-shaped. Additionally, I saw that the inflammatory cells take no place in the fifth cluster.

For the train images, when the k = 5, binNumber = 9, d = 5, and N = 10; I observed that the cell ratios in the first cluster are ranked as epithelial, spindle-shaped, and inflammatory; where the inflammatory cells take no place. I saw an identical relationship also in Cluster 2; however, I observed that the inflammatory cells take a little place. In Cluster 3 for the train images, I saw that whereas the inflammatory cells have the biggest proportion with domination, the epithelial cells have the least proportion. In Cluster 4, I concluded that the ratios of cells are ranked as spindle, inflammatory, and epithelial. In Cluster 5, I saw that the ratios of cells are ranked as spindle, epithelial, and inflammatory; where the inflammatory cells take no place.

Lastly, for the test images, when the k = 5, binNumber = 9, d = 5, and N = 10; I saw that the first cluster is dominated by the spindle-shaped cells. Furthermore, I also observed that the inflammatory cells took no place in Cluster 1. For Cluster 2, I observed that the inflammatory cells significantly dominated the cluster, while the other types of cells took almost no place. In the third cluster, I inferred that while the spindle-shaped cells took the biggest proportion, the inflammatory cells took the smallest. In Cluster 4, I inferred that while the inflammatory cells took the biggest proportion, the epithelial cells took the least. In the fifth cluster, I observed that the spindle-shaped cells took slightly more place than the epithelial cells, and inflammatory cells took almost no place.

In this homework, I applied a row-wise feature normalization where each row corresponds to a cell. For normalizing the features, firstly, I get the mean of the features of each cell by using the np.mean() function coming from the NumPy library. Then, I subtracted this mean value from the value of each cell feature that exists in the feature vector of that cell. Finally, I divided the result of each cell feature by the standard deviation of the features of that cell. I obtained the standard deviation by using the np.std() function coming from the NumPy library. In short, I applied a row-wise z-score normalization for the features that I extracted. While researching, I also found out that this normalization can be performed by using necessary functions under the Standard Scaler module of the sklearn library of Python. Normalizing the features is significantly important for the k-means clustering

algorithm. Because it can make the scales of each cell feature equal and remove the bias in the clustering results. Therefore, by applying feature normalization in k-means, we can get more accurate and interpretable clustering results at the end.

When I checked the total number of spindle-shaped, epithelial, and inflammatory cells in each image, the ratio of each cell type in each cluster, and the visual clustering results; I realized that there exists a class imbalance issue. Because I saw that while one or more cell types have significantly fewer occurrences in an image or cluster, the other cell type(s) dominated that image or cluster. As a result of the class imbalance problem in the images, I observed that some cell types were underrepresented while others were overrepresented by the k-means clustering algorithm. Therefore, I saw that the visual clustering results were not adequately accurate and representative. In order to prevent this class imbalance problem, after researching, I found out that undersampling, weighted k-means clustering algorithm, and oversampling can be applied. In undersampling, the main purpose is to decrease the dominance of the major class by choosing a random set of instances from the major class and removing them from this class. In oversampling, the primary aim is to increase the existence and representation amount of the instances of underrepresented classes by adding new instances to these classes. In the weighted k-means approach, we can assign lower weights to the more dominant classes and higher weights to the underrepresented classes, and therefore prevent the class imbalance to a certain extent.

BONUS PART

For k = 5:

numberOfOrientations = 5, frequency = 0.70, theta = [0, 45, 90, 135, 180], sigma_x = [1, 3], sigma_y = [1, 3], binNumber = 6, d = 5, N = 10

Ratios of Inflammatory, Epithelial, and Spindle-Shaped Cells for Each Cluster (for train images):

	Inflammatory	Epithelial	Spindle-shaped
Cluster 1	0.00	0.2631578947368421	0.7368421052631579
Cluster 2	0.06775067750677506	0.5474254742547425	0.38482384823848237
Cluster 3	0.7744360902255639	0.03759398496240601	0.18796992481203006
Cluster 4	0.004149377593360996	0.5186721991701245	0.47717842323651455
Cluster 5	0.38235294117647056	0.19327731092436976	0.42436974789915966

Ratios of Inflammatory, Epithelial, and Spindle-Shaped Cells for Each Cluster (for test images):

	Inflammatory	Epithelial	Spindle-shaped
Cluster 1	0.6521739130434783	0.0966183574879227	0.25120772946859904
Cluster 2	0.008016032064128256	0.4749498997995992	0.5170340681362725
Cluster 3	0.13256484149855907	0.2824207492795389	0.5850144092219021
Cluster 4	0.00	0.44537815126050423	0.5546218487394958
Cluster 5	0.9058823529411765	0.029411764705882353	0.06470588235294118

Visuals of The Clustering Results:

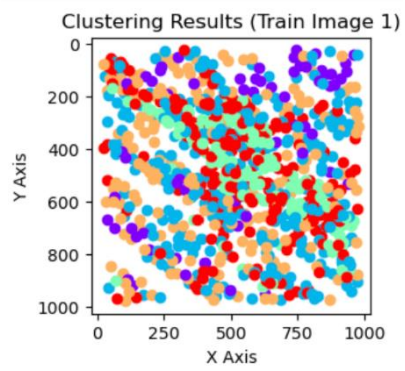

Figure 25: train_8.png Clustering Results

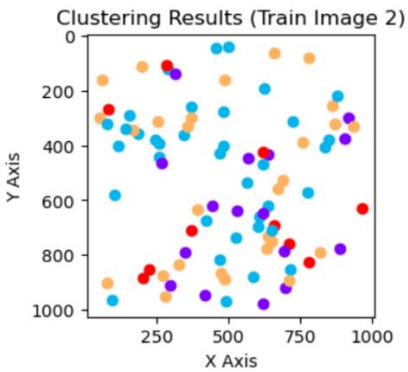

Figure 26: train_11.png Clustering Results

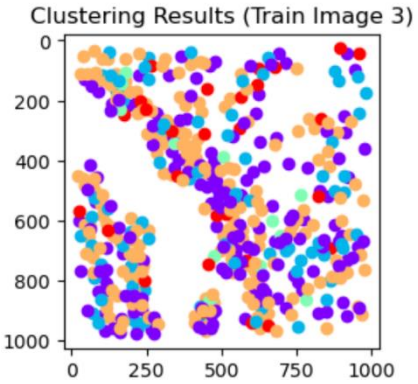

Figure 27: train_14.png Clustering Results

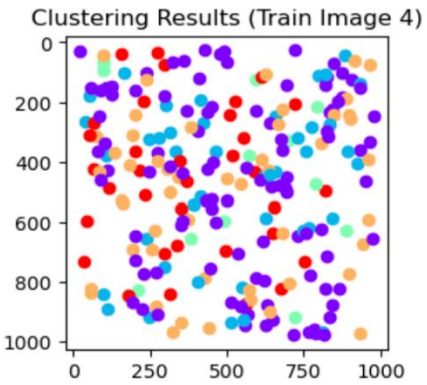

Figure 28: train_21.png Clustering Results

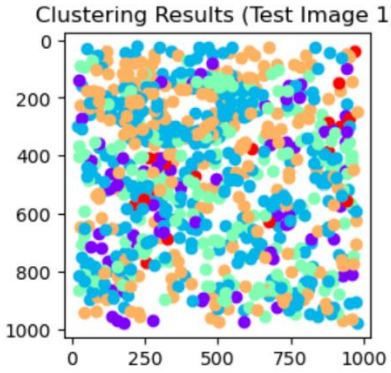

Figure 29: test_1.png Clustering Results

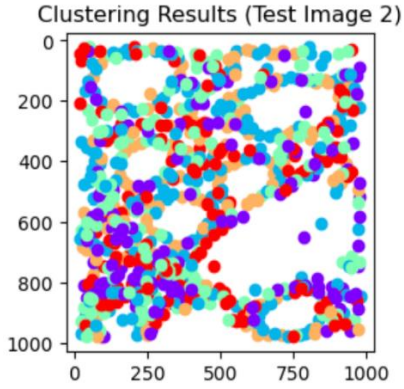

Figure 30: test_10.png Clustering Results

EXPLANATION AND INTERPRETATION OF THE EXPERIMENT RESULTS

In the bonus part, for extracting the textural features, I used a Gabor filter instead of a cooccurrence matrix. I have normalized the features by utilizing the approach I used in Part 2. As the parameter values, I have used 5 for the number of orientations parameter, 0.70 for the frequency parameter, [0, 45, 90, 135, 180] for the theta parameter, [1, 3] for the sigma_x parameter, [1, 3] for the sigma_y parameter, 6 for the binNumber parameter, 5 for the d parameter, and 10 for the N parameter. I have selected the values of these parameters such that they keep the silhouette score as close as possible to 1 and the inertia score as low as possible. Because by their definitions, if the inertia score is low and the silhouette score is closer to 1, then it means that the clustering results are likely to be more consistent, interpretable, and representative. In addition; while selecting the values of these parameters, I tried to obtain compact, accurate, and well-separated clusters in the cluster visualization part. In terms of these aspects, the parameter values which I have used gave me better results. In this question, I conducted the experiments for k = 5. Overall, I obtained unequal distribution of cell types in each cluster and highly imbalanced clustering results for the parameter combination, k value, and train/test images I used. When the images are train images, for Cluster 1, I observed that the spindle-shaped cells dominated the cluster. I also observed that there are no inflammatory cells in this cluster. For Cluster 2, I observed that the ratios of the cells are ranked as epithelial, spindle-shaped, and inflammatory. For Cluster 3, I observed that the ratios of the cells are ranked as inflammatory, spindle-shaped, and epithelial. I also saw that the inflammatory cells dominated the third cluster. For Cluster 4, I concluded that the inflammatory cells took almost no proportion, while the epithelial cells took more than half of the place. In Cluster 4, I also inferred that the spindle-shaped cells have a cell ratio that falls between the ratios of other cell types. For Cluster 5, I saw that the ratios of cells are ranked as spindle-shaped, inflammatory, and epithelial. When the images are test images, for Cluster 1, I observed that the inflammatory cells have the biggest proportion, while the epithelial cells have the smallest proportion. For Cluster 2, I inferred that the inflammatory cells take a minimal place, while the spindle-shaped cells take the most place. I also observed that the ratio of epithelial cells in the second cluster falls between the ratios of other cell types in the Cluster 2. For Cluster 3, I observed that the cell ratios are ranked as spindle-shaped, epithelial, and inflammatory; where spindle cells took the most proportion. For Cluster 4, I saw that the

inflammatory cells take no place and the spindle-shaped cells take the most place. I also observed that the epithelial cells take the second most place within the fourth cluster. Lastly, for Cluster 5, I inferred that the inflammatory cells significantly dominated the cluster, while the other cell types take almost no place. In terms of the effectiveness of the extraction of textural features and the values of silhouette & inertia scores, the performance of the Gabor filter was similar to the cooccurrence matrix approach I used. When I checked the total number of inflammatory, epithelial, and spindle-shaped cells in each image, the ratio of each cell type in each cluster, and the visual clustering results; I realized that there exists a class imbalance issue. Because I saw that while one or more cell types have significantly fewer occurrences in an image or cluster, the other cell type(s) dominated that image or cluster. For handling this problem, the techniques which I previously mentioned in my corresponding discussions can be applied (like undersampling).

Note 1: The execution time of my whole code is approximately 15 minutes.

Note 2: For the libraries I used in my code and which do not exist on your computer, you need to install them by using the necessary **"pip install"** command.

Note 3: Depending on the availability of the built-in Python functions I used in the Python version which you are using on your computer, you may need to do necessary updates or upgrades to the Python version of your computer.

Note 4: You can see all outputs in the "ipynb" file I submitted.