GLOMERULUS CLASSIFICATION

Jack Chun Kit Kot
Department of Chemical and Biological Engineering
`ckkot@connect.ust.hk`

Bingxin Huang
Department of Chemical and Biological Engineering
`bhuangao@connect.ust.hk`

Sarah Suet Ling Chow
Department of Chemical and Biological Engineering
`slschow@connect.ust.hk`

1 Introduction

Histological analysis, the examination of processed and stained thin biological tissue under a microscope, is an important standard clinical practice for diagnosing the disease for subsequent treatment administration. However, the visual inspection of histological images is a laborious and time-consuming task even for experienced pathologists; it requires the identification of disease markers through manually scrutinizing the morphological structures of various types of cells within the tissue. To aid pathologists in accomplishing the arduous task, engineers have been constantly probing the possibility of autonomously classifying different types of cells in a histological image. On the other hand, an accurate tissue classifier can help researchers in the biomedical field be more effective when processed with raw data that involves different stain images.

In this study, we will explore the accuracy of classification of two types of histological images generated from mouse kidney tissue - the ones with glomerulus and the ones without. Manifold learning methods will be used to reduce redundant information and to retain the conducive information in the image. In view of the advance of deep learning, this study would also train a classifier, AlexNet with an untreated dataset and preprocessed one by principal component analysis (PCA) and Generative Adversarial Network (GAN). With the aforementioned data reduction method, we hypothesize that the compressed images will be able to preserve the most distinct information of the image for the classification process and also spare the time for training the classifier while keeping a commensurate classification accuracy.

2 Dataset

The dataset consists of a training set of 803 images and a test set of 80 images, as shown in Table 1, for the previously introduced binary classification problem. Each image is a 256x256 histological image of a biological tissue sample stained by Hematoxylin and Eosin (HE) or Masson Trichrome (MT) staining chemicals and captured with a microscope.

Content	Examples
Training set HE-images with Glomerulus	200
Training set HE-images without Glomerulus	200
Training set MT-images with Glomerulus	191
Training set MT-images without Glomerulus	212
Testing set HE-images with Glomerulus	20
Testing set HE-images without Glomerulus	20
Testing set MT-images with Glomerulus	20
Testing set MT-images without Glomerulus	20

Table 1: The content of the dataset

3 Methodology

3.1 Manifold learning

3.1.1 Multidimensional Scaling (MDS)

Multidimensional Scaling (MDS) [1] is a means of visualizing the level of similarity of individual cases of a dataset. In MDS with classical scaling, the inputs are projected into the subspace that best preserves their pairwise squared distance or, as done in practice, their dot products.

3.1.2 Isometric Mapping (ISOMAP)

Isometric Mapping (ISOMAP) [2] is a nonlinear dimensionality reduction method, and extends metric multidimensional scaling (MDS) by incorporating the geodesic distances imposed by a weighted graph. The ISOMAP algorithm is as follows:

- Determine the neighbors of each point. It contains all points in some fixed radius and K nearest neighbors.
- Construct a neighborhood graph. Each point is connected to another if it is a K nearest neighbor. Edge length equal to Euclidean distance.
- Compute the shortest path between two nodes.
- Compute lower-dimensional embedding.

3.1.3 Locally Linear Embedding (LLE)

Locally linear embedding (LLE) [3] seeks a lower-dimensional projection of the data which preserves distances within local neighborhoods. It is analogous to implementing a series of local Principal Component Analyses yet globally compared for the best non-linear embedding. While other methods of dimensionality reduction fail to preserve local geometric properties on non-linear space, LLE takes advantage of the local geometry and pieces it together to perform and preserve the global geometry on a lower dimensional space. The logic flow of the algorithm is presented below:

- Compute the K nearest neighbors
- Solve for a set of weights that can be used to reconstruct a point
- Compute embedding coordinates using the weights

3.2 Classification by AlexNet

3.2.1 AlexNet

Consider the current rise of deep learning in the field of image processing, it is relatively common for researchers to perform image classification by training a classifier. Among all the classifiers that were Innovated by the rapid development of deep learning, AlexNet[5] as a convolutional neural network architecture that got the first place in 2015 ImageNet contact aroused our interest. Based on its excellent image classification ability, this project proposes such architecture as the target classifier. The structure of the proposed model is demonstrated in figure 1. To enhance the performance of the suggested classifier, it is suggested to adopt data preprocessing by either simplifying the data or extending the data amount to allow training of the proposed network in a more efficient and precise way.

3.2.2 Data Preprocessing

Data Extension with Flipping The effectiveness of the learning process of a classifier usually depends on the amount of data within a dataset. Since the higher the data, the more information can be provided to the model, it is usually preferred to have a dataset with a large amount of data. Flipping original images to extend the given dataset is usually a common method. Yet, consider the adjustment is usually minor, the provided message could be very similar which might lead to the problem of overfitting. Thus, data extension with flipping would be considered one of the baselines on evaluating the classifying ability of the classifier.

Data Reconstruction with PCA Principal Component Analysis (PCA) [6] popular data dimensionality reduction method which transforms a set of correlated variables into a smaller number of uncorrelated variables, namely principal components while preserving the variability of the original data as much as possible. PCA processes original data linearly through calculating and subsequently choosing the eigenvectors with larger eigenvalues to effectively remove
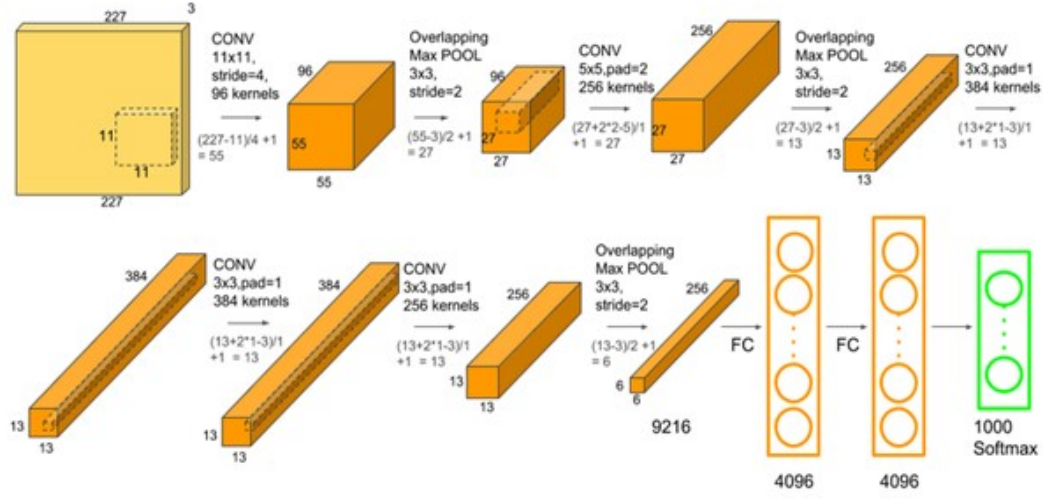
Figure 1: AlexNet Architecture [4]

redundant information in the data set. Besides classifying data into different populations through visualizing data points in a scatter plot, PCA shows immense application into image compression through minimizing the sizes in bytes while keeping the quality of the image. Thus it is chosen to be an image preprocessing method since we hypothesize that PCA would only banish non-essential information in an image and would not affect the classification accuracy drastically.

- All samples were centralized. The average value of the corresponding features of all samples is subtracted from the different features of each sample in the data set, and the average value of the data on the different features processed is 0. The advantage of this method is that it can reduce the difference between features, make different features have the same scale, and make different features have the same influence on parameters.

$$U = \begin{bmatrix} \overline{X}_1 & \overline{X}_2 & ... & \overline{X}_{j-1} & \overline{X}_j \\ \overline{X}_1 & \overline{X}_2 & ... & \overline{X}_{j-1} & \overline{X}_j \\ & \cdot & & & \cdot \\ & \cdot & & & \cdot \\ \overline{X}_1 & \overline{X}_2 & ... & \overline{X}_{j-1} & \overline{X}_j \\ \overline{X}_1 & \overline{X}_2 & ... & \overline{X}_{j-1} & \overline{X}_j \end{bmatrix} \quad (j \, \epsilon \, cols, \quad \overline{X}_j = \frac{\sum_{i=0}^{ierows} x_{(i,j)}}{rows})$$

$$X^* = X - U = \begin{bmatrix} X_{(0,0)} - \overline{X}_1 & X_{(0,1)} - \overline{X}_2 & ... & X_{(0,j-1)} - \overline{X}_{j-1} & X_{(0,j)} - \overline{X}_j \\ X_{(1,0)} - \overline{X}_1 & X_{(1,1)} - \overline{X}_2 & ... & X_{(1,j-1)} - \overline{X}_{j-1} & X_{(1,j)} - \overline{X}_j \\ & \cdot & & & \\ & \cdot & & & \\ X_{(i-1,0)} - \overline{X}_1 & X_{(i-1,1)} - \overline{X}_2 & ... & X_{(i-1,j-1)} - \overline{X}_{j-1} & X_{(i-1,j)} - \overline{X}_j \\ X_{(i,0)} - \overline{X}_1 & X_{(i,1)} - \overline{X}_2 & ... & X_{(i,j-1)} - \overline{X}_{j-1} & X_{(i,j)} - \overline{X}_j \end{bmatrix} \quad (i \, \epsilon \, rows, j \, \epsilon \, cols)$$

Figure 2: PCA

- Calculate the covariance matrix of samples (each column represents a feature and each row represents a sample).
- Eigenvalue decomposition of the covariance matrix is performed to obtain eigenvalues and eigenvectors.

- The projection matrix W is constructed by extracting the eigenvectors corresponding to the largest k eigenvalues.

- For each sample in the sample set, the projection matrix W is multiplied to obtain dimensionality reduction data.

Data Extension with DCGAN   Since capturing a histological image involves time-consuming procedures and requires skilled personnel to process and stain the sample tissue, training data of histological images often encounters a problem of insufficient supply. Given the rise of Generative Adversarial Networks(GAN) [7] advancing the field of image processing, extending the image dataset by synthesizing new images with a trained GAN could be a possible solution to the problem of inadequate image data. As one of the most representative networks in deep learning, GAN follows a deep learning basic framework. Train preprocessed data with a model to predict output until satisfied. Satisfaction of the predicted output based on the performance of loss function, a judgemental rule in the world of deep learning. Superior in the field of image transformation, GAN consists of a generator generating a fake image, and a discriminator identifying whether the input image is real or generated. Different loss functions are applied to measure the difference and similarity of real images and generated fake images and hence optimize both the network until the generated images are nearly indistinguishable from the real one. In terms of image synthesis, DCGAN is one of the representatives. As a model based on GAN, it consists of a pair generator and discriminator. Through input noise along with random input from the dataset, both generator and discriminator of such GAN can learn a hierarchy of representations from object parts to scenes [8]. The trained generator can also be used to synthesize new images of desired classes. Consider its potential of generating distinguishable images, this project will adopt the captioned model as a proposed method that enhances the classifier by extending the dataset with synthesized images. Figure 4 provides a demonstration of the process of generating new images of HE with glomerulus class.
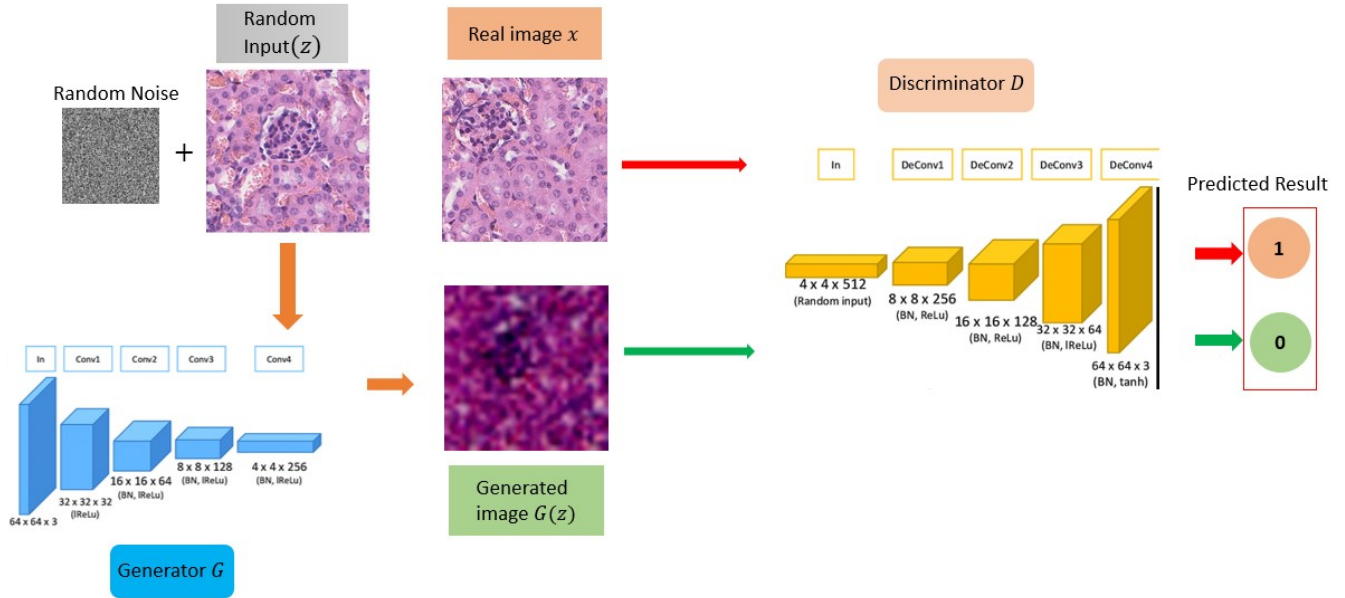

Figure 3: DCGAN Network

Application of DCGAN on PCA data   Despite DCGAN and PCA preprocessing the data from different aspects, the combination of the two approaches might be a potential alternative to train the classifier more effectively in terms of efficiency and accuracy. Considering the simplified data might allow the network to understand the difference between each class clearer, increasing the amount of such simplified data might further accelerate the learning process. Thus, it becomes one of our approaches. To adopt such an application, we performed PCA on all the training data and used the reconstructed data to train DCGAN with 50 epochs. Combined with the original reconstructed training data, generated images by the trained DCGAN model will be used to train the classifier and compare its "Accuracy vs Epochs" curve to the above-mentioned methods.

4

4 Result

4.1 Graph of manifold learning classification results

We visualize the data distribution on the first two principal components with three different manifold learning methods: MDS, ISOMAP, LLE.
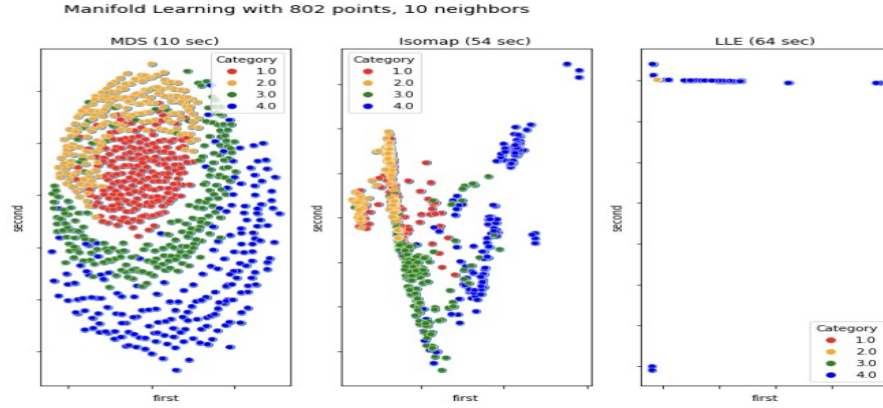

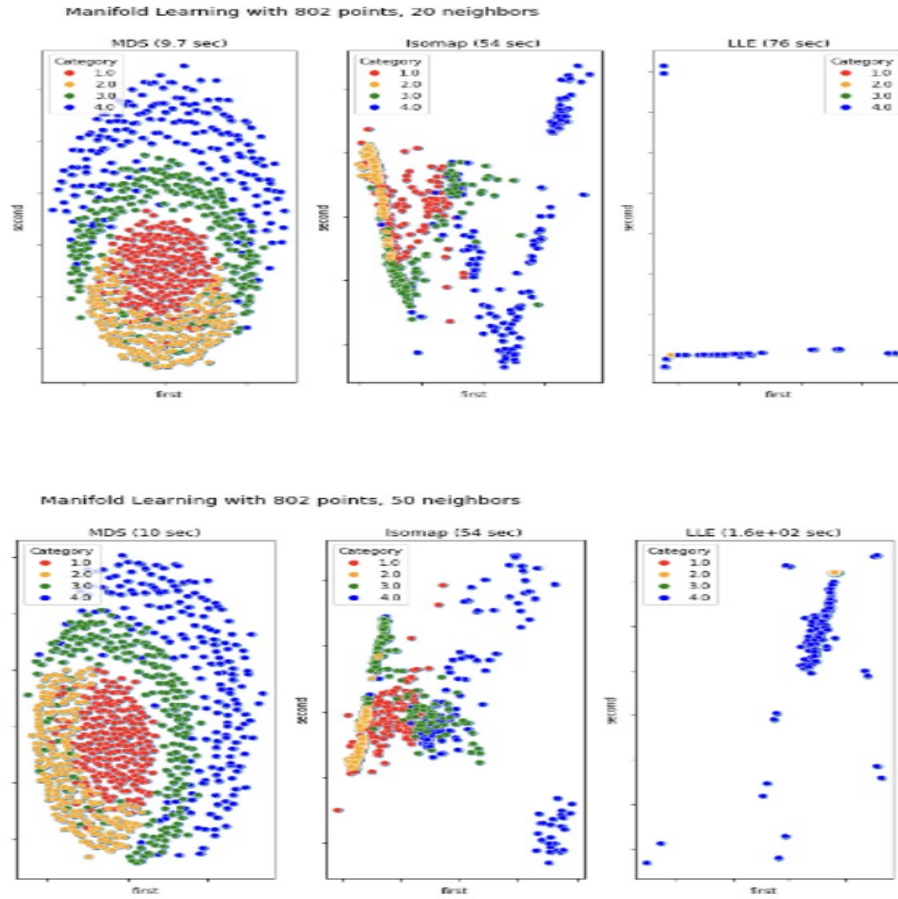Figure 4: Manifold Learning with 802 points, 10 neighbors

Figure 5: Manifold Learning with 802 points, 20 and 50 neighbors

Figure 4 (a) shows the MDS visualization result and different classes represented by different colors. From the figure, we notice that dots of different colors separate well, which means images with the same features group together and the cluster separates well with other classes. On the other hand, the red dots (HE with glomerulus) and the yellow dots (HE without glomerulus) seem to be clusters due to the similarity of their staining method. It shows that the MDS method has a good performance on visualization and classification in this case.

Figure 4 (b) shows the visualization result by ISOMAP. The different dots scatter and mix to some extent, but can still distinguish between four classes. For example, the green dots (MT without glomerulus) and the blue dots (MT with glomerulus) are mixed but separate from others.

Figure 4 (c) shows the visualization result by LLE. It seems that most of the dots are clustered around one line, which is difficult to distinguish between different classes.

In the three methods, MDS shows good performance and fast speed. We also change the number of neighbors in the three methods, as shown in Figure 5. It shows that different neighbor numbers affect the performance of the ISOMAP method a lot while affecting the speed of the LLE method a lot.

4.2 Images processing for classification

4.2.1 PCA reconstruction images

PCA reconstruction has been implemented to all the images, where only 100 principal components (PC) of the images were preserved. The designated number of PCs preserved was obtained empirically with a visual inspection, where the reconstructed image is not corrupted and deviated from the original image. The following shows the PCA reconstruction results with different numbers of PCs specified:
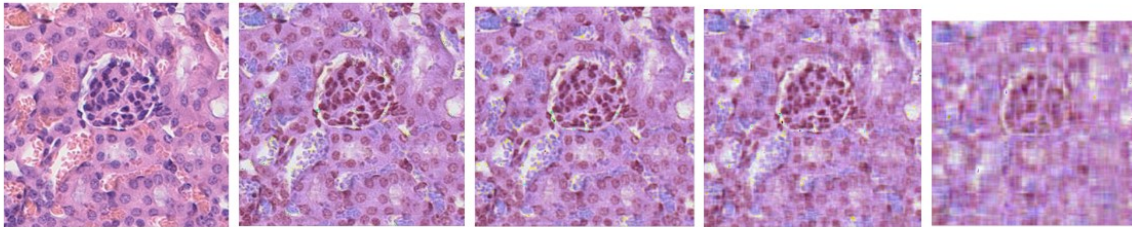

Figure 6: Comparing original picture with PCA reconstructed picture with PC = 100, 60, 30, 10 (left to right)

Comparing the original image with the PCA reconstructed image with only 100 PCs, it is clear that a majority of features of the image are well-preserved, whereas features started to fade out as the number of PCs decreased.

4.2.2 DCGAN generated images

To extend the dataset, A DCGAN was trained for 50 epochs with both original training data and PCA reconstructed data. The trained network was used to generate extra 196 images for each class. The generated image was initially with the size of 32x32 and resized to 256x256. Despite the generated results being a bit blurry, they still preserved the colour tone and captured the specified feature from each class. Examples of the generated image can found in appendix.

4.3 Graphs for Accuracy Comparison by Training Preprocessed Different Datasets

The classifier was trained with four datasets that were preprocessed with different treatment along with the original dataset for 10 epochs. Except for the dataset that is preprocessed only with PCA, other datasets preserve the original data and are extended by the generated or adjusted images. The model was tested for its accuracy every two epochs and the collected data are used to plot the curve of accuracy vs epochs.
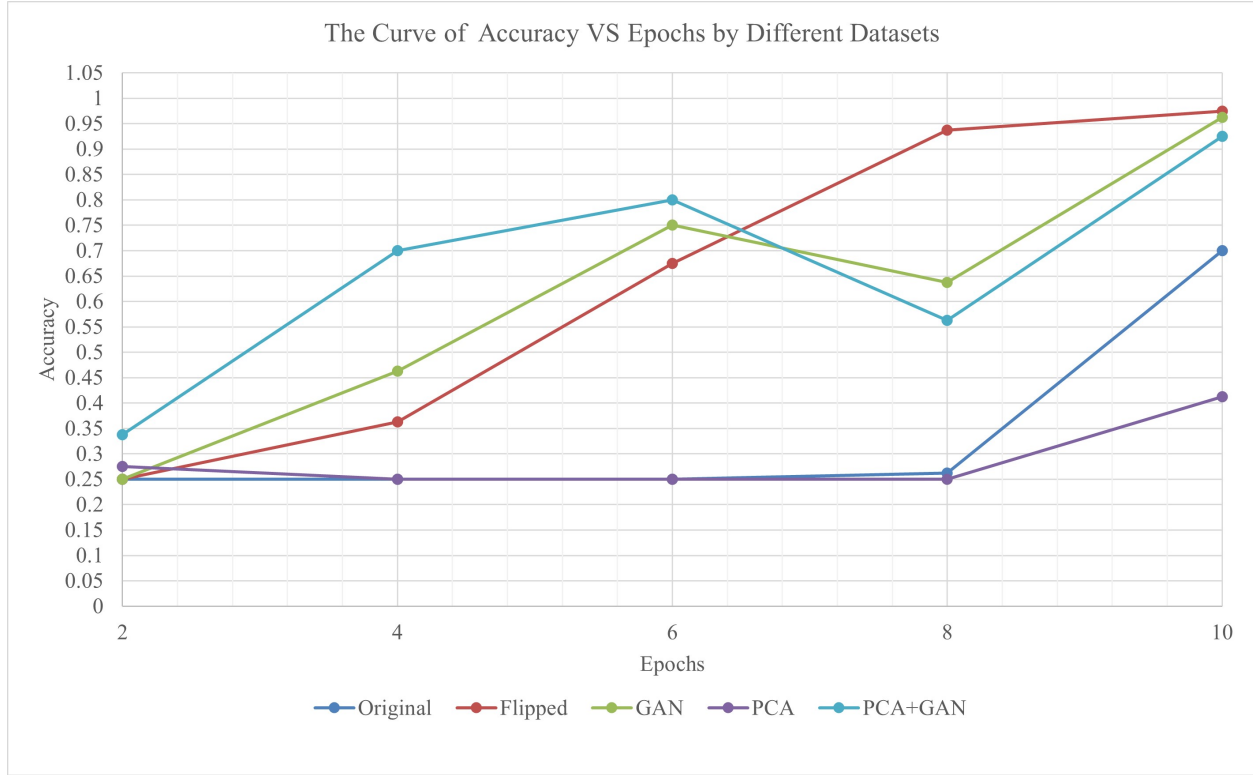
Figure 7: The Curve of Accuracy vs Epochs by Different Preprocessed Datasets

From the obtained results, it could be concluded that the accuracy of the model with any version of datasets increases along with the training epochs. All models have a relatively low accuracy during small epochs. In terms of efficiency, the dataset that was preprocessed with the combination of original and flipping data achieves the highest accuracy and outperforms other methods. Besides, the Data processed with GAN alone also achieves relatively high accuracy. Such observations indicate the preprocessing of data extension can enhance the efficiency of training a classifier. This phenomenon can also be validated from the case of the dataset preprocessed with PCA and GAN. Although the performance of using single PCA preprocessed data did not achieve a high level of accuracy, the application of using GAN to further extend its data indeed can observe an improvement while training the classifier. It also indicates that the more the data within a dataset, the more effective the training of the classifier.

For data preprocessed with PCA, the performance is worse than other proposed methods. It could be attributed to PCA did not preserve the important feature of the data as we hypothesized. Meanwhile, since the amount of original data is relatively low, inadequate data can also be the reason for the unsatisfied performance of the dataset preprocessed with PCA.

On the other hand, despite the performance of the dataset preprocessed with adding flipping images slightly higher than using GAN or a combination of GAN and PCA, such application on extending the data could be quite limited. The amount of flipping data usually depends on the amount of original data while the generation of new images by GAN can be infinite. From the aspect of stability, using GAN to perform data extension might be a better option in terms of enhancing the performance of a classifier.

Compared with classification by manifold learning methods, training a classifier with GAN and PCA preprocessed dataset is superior. Despite manifold learning methods like MDS can classify, they cannot achieve a high accuracy that the classifier can.

5 Conclusion and Future Work

In this project, we tested 3 different methods of manifold learning to visualize all the data. We also utilize Alexnet for the classification of kidney tissues that contain glomerulus in histological images in conjunction with the reconstruction by PCA and generation by DCGAN.

The classification results show that datasets that are processed with data extension achieved higher accuracy than using a simple dimension reduction method. In the future, we hope to apply our method to classify other tissue types such as classifying normal tissue and tumor tissue.
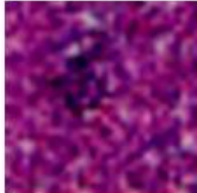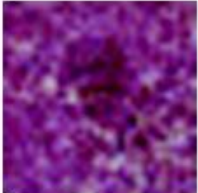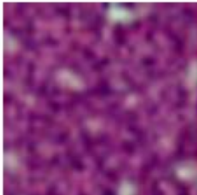
5.1 Contribution

Manifold learning, DCGAN, PCA, AlexNet (All together)

References

[1] I . Borg and P . Groenen. Review reviewed work (s): Modern multidimensional scaling: Theory and applications. *Journal of Educational Measurement*, 3(3):277–280, 2019.

[2] M. Balasubramanian and E. L. Schwartz. The isomap algorithm and topological stability. *Science*, 295(5552):7, 2002. doi: 10.1126/science.295.5552.7a.

[3] X. Xu X. Zhang Y. Hou, P. Zhang and W. Li. Nonlinear dimensionality reduction by locally linear inlaying. *IEEE Trans. Neural Networks*, 20(2):300–315, 2009. doi: 10.1109/TNN.2008.2005582.

[4] Alexnet - imagenet classification with convolutional neural networks. https://neurohive.io/en/popular-networks/alexnet-imagenet-classification-with-deep-convolutional-neural-networks/ (accessed May 20, 2021).

[5] Imagenet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, 2017. doi: 10.1145/3065386.

[6] I. T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016. doi: 10.1109/TNN.2008.2005582.

[7] I. Goodfellow et al. Generative adversarial networks. *Commun. ACM*, 63(11):139–144, 2014. doi: 10.1145/3422622.

[8] Unsupervised representation learning with deep convolutional generative adversarial networks. *4th Int. Conf. Learn. Represent. ICLR 2016 - Conf. Track Proc.*, pages 1–16, 2016.

Appendix

Image type	Original + DCGAN	PCA + DCGAN
HE with glomerulus		
HE without glomerulus		
MT with glomerulus		
MT without glomerulus		