Assess whether the heterogeneity in the medical images affects the clustering of image features extracted from CNN[1]

Author names and affiliations:

Yihan Wang

Research Opportunity Program: STA299Y1

Department of Statistical Science, University of Toronto

Email address: christinee.wang@mail.utoronto.ca

Supervisor: Professor Pascal Tyrrell

Keywords:

Heterogeneity, Cluster analysis, Image features, Convolutional Neural Network, Transfer learning

[1] Acknowledgement:
Thanks for the supervision of Prof.Pascal Tyrrell, and the help from Mauro Mendez and Atsuhiro Hibi.
ChatGPT is used to reword sentences in Abstract, Introduction, Methodology (Data Section) and Discussion.

Abstract

In this study, we assessed whether heterogeneity in medical images could affect the clustering of image features extracted by a convolutional neural network (CNN). We utilized two publicly available datasets and conducted binary classification tasks using the EfficientNetB0 architecture. We then performed cluster analysis on the CNN-extracted image features using K-means clustering algorithm and retrained the binary CNN classifier on the images in each cluster. Our findings suggest that the presence of heterogeneity in medical images can impact the clustering of image features extracted by CNN. We also found a relationship between the performance of the retrained models and the distances between clusters, with closer clusters exhibiting more similar retrained model performances.

Summary Statement: This study found that the presence of heterogeneity in medical images can affect the clustering of image features extracted by CNN, and that the performance of retrained models is related to the distances between clusters, with closer clusters exhibiting more similar retrained model performances.

Introduction

Premise: Machine learning is a technique for recognizing patterns that have been widely applied in medical image analysis [1], one such machine learning model is the convolutional neural network (CNN), which automatically learns the features that are believed to be of importance in making the prediction or diagnosis of interest and extract it for medical image understanding [2]. CNN model has been widely used in medical image classification tasks, involving the extraction of features from the image, and assigning labels using the extracted features. Image features are specific visual patterns or characteristics that are present in an image, such as edges, corners, or textures. However, due to the differences in patient populations and clinical disparities across hospitals, medical imaging data are always heterogeneous [3]. Therefore, when applying the CNN model to classify medical images, it is important to find methods to assess the effect of heterogeneity on CNN model training.

Literature Review: A previous study proposed to use cluster analysis, a statistical method for grouping image-based features with similar characteristics, to assess the impact of dataset

heterogeneity on deep convolutional network accuracy [4], the study tested how the CNN model performs among clusters in the test set. The differences in clusters' accuracy indicated the existence of heterogeneity and its effect on the CNN model generalizability. Our study is based on the previous study to use cluster analysis to assess the impact of heterogeneity on the clustering of image features extracted from the CNN model.

Gap: Research has been conducted in analyzing the effect of heterogeneity on CNN model performance using methods such as cluster analysis [4], it was shown that cluster analysis can be used as a tool to identify if the CNN model is affected by heterogeneity from the medical imaging dataset. However, it remains unknown how the heterogeneity in the medical images could affect the clustering of image features extracted from the CNN model. Therefore, except for comparing the CNN model performances on the clusters in the testing group, we further retrained CNN models using images in each cluster in the training set to assess how heterogeneity could affect the clustering of image features.

Purpose: To assess whether the heterogeneity in the medical images affects the clustering of image features extracted from the CNN model.

Research Question: Will the CNN models retrained from the medical images in each cluster obtained from the clustering of image features extracted from CNN, perform differently?

Hypothesis: The heterogeneity in the medical images will affect the clustering of image features extracted from CNN in the sense that the CNN models retrained from each cluster perform differently.

Objectives:

1. Compare the test accuracy of the overall binary CNN model in the clusters assigned in the testing set.
2. Describe the re-training procedure of the binary CNN models from the medical images in each cluster obtained from the clustering of image features extracted from CNN.

3. Compare the retrained CNN models' performances on the testing set to see whether they differ.
4. Explain how the heterogeneity in the dataset affects the clustering of image features

Methods:

Data:

1. Fundus Images (Cataract vs. Normal) Dataset: We used the Ocular Disease Intelligent Recognition dataset from Kaggle [5], which contains ophthalmic data of 5,000 patients including their age, sex, and color fundus photographs of left and right eyes, along with doctors' diagnostic keywords. The dataset consists of several columns such as "ID," "Patient Age," "Patient Sex," "Left-Fundus," "Right-Fundus," "Left-Diagnostic Keywords," "Right-Diagnostic Keywords," and 8 classes of eye diseases with their corresponding labels. We focused on binary classification between two classes, "cataract" and "normal," and selected fundus images with diagnostic keywords as either "normal fundus" or "cataract." We combined the selected images from left and right eyes to obtain a preprocessed dataset consisting of 1036 rows. This dataset was collected from different hospitals/medical centers in China by Shanggong Medical Technology Co., Ltd., using various cameras in the market, such as Canon, Zeiss, and Kowa, resulting in varied image resolutions. To ensure the model's robust evaluation, we stratified the data based on the two classes, "cataract" and "normal," and split it into training and testing sets in the ratio of 0.8 and 0.2, respectively. We further did data augmentation on the training data and obtain a total of 1428 images for training, the number of images in the testing set is 208.

2. Chest X-Ray (Cardiomegaly vs. Normal) Dataset: We obtained the Chest X-Ray dataset from Kaggle [6], which is a random subset of the NIH Chest X-ray Dataset containing 5,606 images with a size of 1024 x 1024. The dataset has 15 classes that include 14 diseases and one class for "No findings". To perform a binary classification task, we chose "Cardiomegaly" and "No findings" as our target classes. There were a total of 141 images labelled as "Cardiomegaly" and 3044 images labelled as "No findings". To keep the classes balanced, we filtered the normal images to match the size of the cardiomegaly images. We stratified the data based on the outcome (cardiomegaly or normal) and split it into training and testing sets with a proportion of 0.8 vs 0.2, respectively. Since the training set is too small, we further did data augmentation on

the training data and obtain a total of 450 images for training, the number of images in the testing set is 57.

Binary Classification CNN Model: We used the Keras library, a Python-based high-level neural network API, to build and train a binary classification model based on the EfficientNetB0 architecture, a state-of-the-art convolutional neural network that has demonstrated strong performance on image classification tasks. The pre-trained EfficientNetB0 model, which was trained on the large-scale ImageNet dataset, was used as a starting point. To fine-tune the model for our specific binary classification task, we added a custom output layer consisting of a single neuron with a sigmoid activation function. During training, the pre-trained layers were frozen, and only the output layer was trained using the binary cross-entropy loss function and the Adam optimizer. Early stopping and model checkpoint techniques were used to prevent overfitting. We did not use cross-fold validation since we utilized all training data for training the model.
To classify cataract vs. normal (Fundus Images), the model was trained for 10 epochs with a batch size of 8 on the training set, while monitoring its performance on the validation set. EarlyStopping was applied when the validation accuracy did not improve for three consecutive epochs. To classify cardiomegaly vs. normal (Chest X-Ray Images), the model was trained for 30 epochs with a batch size of 8 while monitoring its performance on the validation set. EarlyStopping was applied when the validation accuracy did not improve for 10 consecutive epochs, as this model is more difficult to train than the model to classify cataract and normal.

Extract image features from the medical images: To extract features from the medical images, we removed the last layer of the model and created a new model that outputs the activations of the second last layer. We used this new model as a feature extractor to transform the input images into a lower-dimensional space of extracted features. Both the features in the images of the training set and testing set were extracted by the CNN model.

Cluster analysis on the image features extracted from the CNN: We conducted cluster analysis on image features extracted from a CNN. Since the extracted features have high dimensionality, we used principal component analysis (PCA) to reduce the dimensionality and applied it to the training and testing data of fundus images. We selected the first 200 principal components to

explain over 70% of the variance. For the Chest X-Ray Dataset, the test set only have 57 images, selecting the maximum number of components (57) could only explain 40% variance for the features in the training set, we actually compared two approaches to do the cluster analysis, in the first approach, we do the cluster analysis directly based on the image features extracted from CNN, in the second approach, we do the cluster analysis based on the 57 principal components. We then used the K-means clustering algorithm to group images with similar characteristics into four clusters, following the previous work from [4]. The K-means algorithm is preferred due to its computational efficiency, scalability, and simplicity. We assigned the transformed testing set into the four clusters based on the cluster centers found from the training set. For each cluster assigned in the testing set, we calculated the accuracy of the binary CNN classification model trained above.

Retrain the binary classifiers on each of the clusters in the training set: To retrain the model on the clusters, we first obtained the cluster labels for each sample in the dataset using the K-means algorithm. We then created separate data frames for each cluster containing the image features of the images belonging to the corresponding cluster. Each model was then trained using the images belonging to the corresponding cluster. We used the same CNN architecture and hyperparameters that were used in the initial binary classification model training. However, we only used the samples belonging to the corresponding cluster for training each model.

Train binary classifiers on random medical images with the same size as each cluster: To ensure a fair comparison of the performances of the retrained models from each cluster, we also trained a binary classification model on random medical images from the training set with the same size as each of the four clusters. This is to prevent any bias introduced by the varying sample sizes in each cluster when comparing their performances. We randomly sampled the same number of images as each cluster from the training set, created four separate data frames, and trained a binary classification model on each of them. To obtain more reliable estimates of the model's performance, we trained the model five times for each sample size and averaged the performance metrics over the five runs.

Experiments:

Compare the test accuracy of the CNN model in the clusters assigned in the testing set: For both the Fundus Images Dataset and the Chest X-Ray dataset, we tested the performance of the binary classifier trained from the data in the training set on the four clusters assigned in the testing set and compared the testing accuracy of the model on the test clusters.

Compare the performances of the models retrained from each cluster: For both the Fundus Images Dataset and the Chest X-Ray dataset, the performance metrics (Test Accuracy, AUC, Sensitivity, Specificity) of the models retrained from each cluster were compared. As the cluster sizes vary, we also compared the retrained models with the performance of models trained from random medical images with training size equivalent to their corresponding cluster.

Statistics: The main outcome measure in our study is the model testing accuracy, which is the proportion of the total number of corrected classified samples among the total number of samples in the testing set. Additionally, AUC, Specificity and Sensitivity will also be taken into account when comparing the model performances. We measured the between-cluster distance to observe how distant clusters are from each other. The between-cluster distance is calculated based on the distances between the centroids of each cluster. We used intra-cluster distance to measure the similarity of the images in each cluster, the intra-cluster distance is a measure of how close the data points are to each other within a cluster. It is calculated as the average distance between all pairs of data points within the cluster. Clusters with smaller intra-cluster distances are considered to be more compact and homogeneous, while clusters with larger intra-cluster distances are less compact and less homogeneous [7].

Environment: The experiments are done using Google Colab, and the library used for training the CNN model is Keras.

Results:

Table 1: The CNN model accuracy on the clusters in the test set. [Fundus Images Dataset]

Accuracy on Whole test set	Test Cluster1 Accuracy	Test Cluster2 Accuracy	Test Cluster3 Accuracy	Test Cluster4 Accuracy
0.9712	1	0.97	0.9694	1

Table2: The CNN model accuracy on the clusters in the test set. [Chest X-Ray Dataset]

Accuracy on Whole test set	Test Cluster1 Accuracy	Test Cluster2 Accuracy	Test Cluster3 Accuracy	Test Cluster4 Accuracy
0.8246	0.9231	1.0	0.8235	0.7351

From Table1 & 2, we could see that the overall binary CNN classification model perform differently on the clusters in the test set for both the Fundus Images Dataset and Chest X-Ray Dataset.

Table3: The performance of the CNN models retrained from the clusters [Fundus Images dataset]

Cluster	Size(n)	Intra-Cluster Distance	Cataract vs. Normal	Retrained Model Test Accuracy	Random Sampled Model Test Accuracy	Retrained Model AUC	Random Sampled Model AUC	Retrained Model Sensitivity	Random Sampled Model Sensitivity	Retrained Model Specificity	Random Sampled Model Specificity
0	42	335	42 vs. 0	50.00%	93.4%	50%	97%	100.00%	92.11%	0	94.62%
1	441	195	20 vs. 421	81.73%	94.23%	96%	98%	67.31%	92.11%	96.15%	96.35%
2	274	428	144 vs. 130	93.75%	94.13%	97%	98%	92.31%	91.15%	95.19%	97.12%
3	671	201	622 vs. 49	95.67%	95.48%	97%	98%	94.23%	93.65%	97.12%	97.31%

Figure1: Distances between the clusters in the fundus image dataset

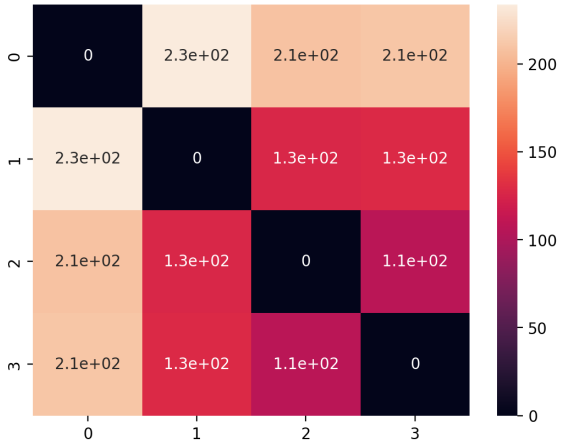


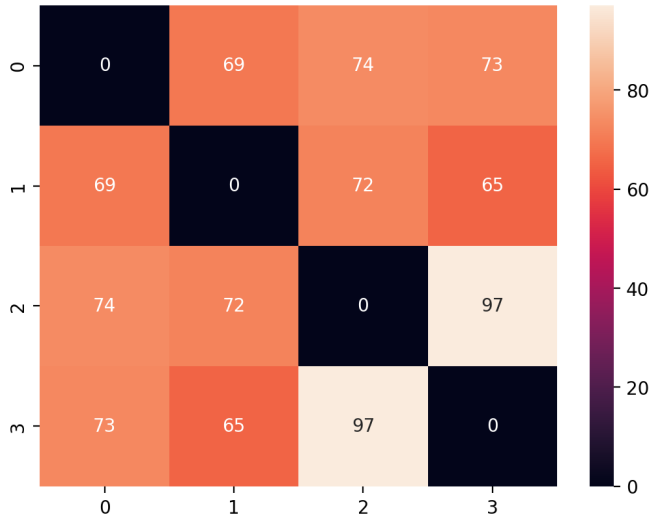
Table 3 reveals that there are two large clusters (Cluster 1 and Cluster 3) in the fundus image dataset. Most images in Cluster 1 are normal, while most images in Cluster 3 are cataracts. The intra-cluster distance of these two clusters is much smaller than the other two clusters, indicating that the images within each of these two clusters are more similar. The retrained models from Clusters 1, 2 and 3 exhibit good performances, while the retrained model from Cluster 0 only has

a test accuracy of 50%. Figure 1 shows that Clusters 1, 2, and 3 are close to each other, with Clusters 2 and 3 being the closest, while Cluster 0 is distant from the other three clusters. Noted that the retrained model from Cluster 0 has 100% sensitivity and 0 specificity, indicating that this model has learned significant features to identify cataracts but has failed to identify normal images. Also, the sensitivity of the retrained model from Cluster 1 is low, suggesting that the images in Cluster 1 may lack crucial features for the model to learn to identify cataracts.

Table 4: The performance of the CNN models retrained from the clusters [Chest X-Ray dataset]

Cluster	Size(n)	Intra-Cluster Distance	Cardiomegaly vs. Normal	Retrained Model Test Accuracy	Random Sampled Model Test Accuracy	Retrained Model AUC	Random Sampled Model AUC	Retrained Model Sensitivity	Random Sampled Model Sensitivity	Retrained Model Specificity	Random Sampled Model Specificity
0	96	239	42 vs. 54	73.68%	70.87%	75%	77%	51.72%	67.59%	96.43%	74.28%
1	108	212	42 vs. 66	80.70%	75.08%	80%	80%	89.66%	78.62%	71.43%	71.43%
2	113	224	53 vs. 60	77.19%	74.03%	79%	78%	89.66%	78.62%	64.29%	69.29%
3	133	238	87 vs. 46	56.14%	73.3%	67%	79%	96.55%	73.1%	14.29%	73.57%

Figure2: Distances between the clusters in the Chest X-Ray dataset



According to Table 4, the sizes of the four clusters are similar. The retrained models from Clusters 0, 1, and 2 exhibit similar test accuracies, with the model from Cluster 1 performing the best and having the smallest intra-cluster distance. Figure 2 shows the distances between the clusters in the Chest X-Ray dataset, revealing that Clusters 2 and 3 are distant from one another, and the retrained models from these two clusters exhibit markedly different performances. Noted

that among the four clusters, the retrained model from Cluster 0 has the highest specificity and lowest sensitivity, indicating that the images in this cluster possess crucial features for the model to recognize normal images but lack vital features to identify cardiomegaly images. Conversely, the retrained model from Cluster 3 has the highest sensitivity and lowest specificity, implying that the images in Cluster 3 have important features for the model to learn to identify cardiomegaly images but lack important features for the model to learn to identify normal images.

Discussion:

The results in Tables 1 and 2 indicate that the binary CNN classification model's overall performance varies across clusters in both the Fundus Images and Chest X-Ray datasets. This finding suggests that heterogeneity in the data can impact the model's generalizability. The results are consistent with the previous study [4] that clustering analysis can be used in a way to show the impact of heterogeneity on the CNN model. However, in the Fundus Images dataset, we only observed a minor difference in the clusters' testing accuracies. One possible explanation for this outcome is that the binary CNN classification model can easily identify cataract or normal fundus images, leading to consistently high accuracy.

The retrained models from the four clusters in the Fundus Image dataset performed differently, which indicates the existence of heterogeneity could affect the clustering of images features. Further analysis revealed that the three clusters with similar good retrained model performances are also spatially close to each other, whereas the cluster with the lowest performance is situated farther away from them. This suggests that there may be a relationship between the performance of the retrained models and the distances between the clusters, with closer clusters exhibiting more similar retrained model performances. The intra-cluster distance of the two largest clusters in the fundus image dataset were found to be much smaller than the other two clusters, indicating that the images within each of the two clusters are more similar. Additionally, one of these two largest clusters mainly consists of normal images, while the other is primarily composed of cataract images. However, there is a cluster which only comprises cataract images. This is likely due to the difference in the image scales in this cluster compared to the other clusters, making it

an outlier in the dataset. Based on the distribution of the four clusters in the Fudus image dataset, it appears that there are mainly two clusters each represent one of the two classes.

The Chest X-Ray dataset also demonstrates varying performances of retrained models across the four clusters. We observed that the three clusters with similar retrained model performances are spatially close to each other, while the cluster with the lowest performance is further away from one of the three clusters but closer to the other two. The retrained model from the lowest performing cluster has high sensitivity but low specificity, indicating that the images in the cluster have important features for identifying cardiomegaly but lack features for classifying normal images. Furthermore, we found that the intra-cluster distance of the two clusters with the best retrained model performances is smaller than the other two clusters. However, we must exercise caution in interpreting this result as it is not always the case that a smaller intra-cluster distance leads to better retrained model performance. We also need to consider whether the images in the cluster are balanced. It is also important to note that even though the retrained models from some of the clusters have good performance, they may not be suitable for practical use. This is because the models are trained using only the images in their corresponding cluster, which are homogeneous and may lack generalizability to other groups.

Limitation:

The small size of the test set (57 images) in the Chest X-Ray dataset limits the ability to capture a representative range of features, with the maximum number of components (57) only explaining 40% of the variance in the training set. To address this, we compared two approaches for cluster analysis: one directly based on CNN-extracted image features, and the other using 57 principal components. While the second approach could not provide meaningful clusters, the first approach (without PCA) yielded more meaningful results and was therefore used for further analysis. However, the large dimensions of the image features may affect the K-means algorithm as well, further research should be done using 4-fold cross-validation to make sure the test set is not too small.

Conclusion:

This study demonstrates that the presence of heterogeneity in medical images can have an impact on the clustering of image features extracted from CNN, as evidenced by the differences in performance among the retrained models from the clusters. Moreover, we found that there may exist a relationship between the performance of the retrained models and the distances between the clusters, with closer clusters exhibiting more similar retrained model performances. We also observed that there are mainly two clusters each representing one of the two classes. Further research can expand the sample size to assess whether it will eventually lead to two large clusters, each representing one class, while the images in the other smaller clusters may contain outlier images in terms of scale and size.

References:

[1] B. J. Erickson, P. Korfiatis, Z. Akkus, and T. L. Kline, “Machine Learning for Medical Imaging,” Radiographics, vol. 37, no. 2, pp. 505–515, Mar. 2017, doi: https://doi.org/10.1148/rg.2017160130

[2] D. R. Sarvamangala and R. V. Kulkarni, “Convolutional neural networks in medical image understanding: a survey,” Evol Intell, vol. 15, no. 1, pp. 1–22, Mar. 2022, doi: https://doi.org/10.1007/s12065-020-00540-3

[3] M. J. Willemink et al., “Preparing medical imaging data for machine learning,” Radiology, vol. 295, no. 1, pp. 4–15, Feb. 2020, doi: https://doi.org/10.1148/radiol.2020192224

[4] M. Mendez, S. Calderon, and P. N. Tyrrell, “Using Cluster Analysis to Assess the Impact of Dataset Heterogeneity on Deep Convolutional Network Accuracy: A First Glance,” Communications in Computer and Information Science, vol. 1087 CCIS, pp. 307–319, 2020, doi: https://doi.org/10.1007/978-3-030-41005-6_21

[5] “Ocular Disease Recognition | Kaggle.” https://www.kaggle.com/datasets/andrewmvd/ocular-disease-recognition-odir5k

[6]“Random Sample of NIH Chest X-ray Dataset | Kaggle.” https://www.kaggle.com/datasets/nih-chest-xrays/sample

[7] A. Binu Jose and P. Das, "A Multi-objective Approach for Inter-cluster and Intra-cluster Distance Analysis for Numeric Data," in Soft Computing: Theories and Applications, R. Kumar, C. W. Ahn, T. K. Sharma, O. P. Verma, and A. Agarwal, Eds., vol. 425, Singapore: Springer, 2022, pp. 401-410, https://doi.org/10.1007/978-981-19-0707-4_30

Tables:

Table 1: The CNN model accuracy on the clusters in the test set. [Fundus Images Dataset]

Accuracy on Whole test set	Test Cluster1 Accuracy	Test Cluster2 Accuracy	Test Cluster3 Accuracy	Test Cluster4 Accuracy
0.9712	1	0.97	0.9694	1

Table 2: The CNN model accuracy on the clusters in the test set. [Chest X-Ray Dataset]

Accuracy on Whole test set	Test Cluster1 Accuracy	Test Cluster2 Accuracy	Test Cluster3 Accuracy	Test Cluster4 Accuracy
0.8246	0.9231	1.0	0.8235	0.7351

Table 3: The performance of the CNN models retrained from the clusters [Fundus Images dataset]

Cluster	Size(n)	Intra-Cluster Distance	Cataract vs. Normal	Retrained Model Test Accuracy	Random Sampled Model Test Accuracy	Retrained Model AUC	Random Sampled Model AUC	Retrained Model Sensitivity	Random Sampled Model Sensitivity	Retrained Model Specificity	Random Sampled Model Specificity
0	42	335	42 vs. 0	50.00%	93.4%	50%	97%	100.00%	92.11%	0	94.62%
1	441	195	20 vs. 421	81.73%	94.23%	96%	98%	67.31%	92.11%	96.15%	96.35%
2	274	428	144 vs. 130	93.75%	94.13%	97%	98%	92.31%	91.15%	95.19%	97.12%
3	671	201	622 vs. 49	95.67%	95.48%	97%	98%	94.23%	93.65%	97.12%	97.31%

Table 4: The performance of the CNN models retrained from the clusters [Chest X-Ray dataset]

Cluster	Size(n)	Intra-Cluster Distance	Cardiomegaly vs. Normal	Retrained Model Test Accuracy	Random Sampled Model Test Accuracy	Retrained Model AUC	Random Sampled Model AUC	Retrained Model Sensitivity	Random Sampled Model Sensitivity	Retrained Model Specificity	Random Sampled Model Specificity
0	96	239	42 vs. 54	73.68%	70.87%	75%	77%	51.72%	67.59%	96.43%	74.28%
1	108	212	42 vs. 66	80.70%	75.08%	80%	80%	89.66%	78.62%	71.43%	71.43%
2	113	224	53 vs. 60	77.19%	74.03%	79%	78%	89.66%	78.62%	64.29%	69.29%
3	133	238	87 vs. 46	56.14%	73.3%	67%	79%	96.55%	73.1%	14.29%	73.57%

Figures:

Figure 1: Distances between the clusters in the fundus image dataset

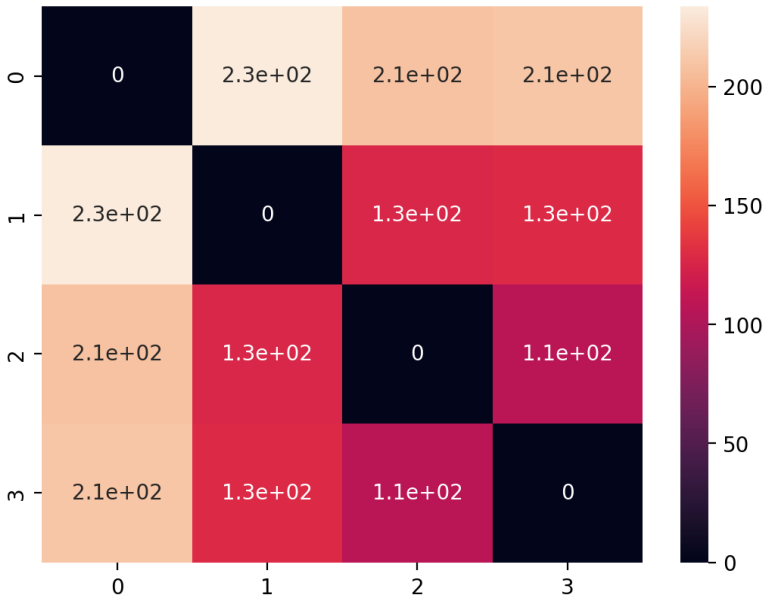


Figure 2: Distances between the clusters in the Chest X-Ray dataset