PaletteMatch

Your personalized art recommendation system

Introduction

Literature Review

Problem Definition

Results and Discussion

Methods

References

Name	Project Proposal	Midterm Report	Final Presentation
Siddharth Chilluvuri	Project Research Proposal Write-Up Project Webpage Creation Video Recording	Worked on implementing KMeans method Implemented a quantitative metric Assisted with midterm report	Worked on Random Forest method and assisted in written deliverables as well as recording the video
Shaktik Bhattacharrya	Proposal Write-Up Project Research	Worked on data preprocessing Worked on putting together and updating midterm report	Worked on preprocessing methods and assisted in written deliverables as well as recording the video
Pranav Murthy	Slideshow Creation Project Research	Worked on data preprocessing Worked on creating visualizations	Worked on new preprocessing methods as well as setting up PACE ICE and assisted in written deliverables
Shadi Raja Buchanan	Gantt Chart Project Webpage Creation Video Recording	Worked on data preprocessing Worked on creating visualizations	Worked on GMM method and assisted in written deliverables as well as recording the video
Cole McCord	Slideshow Creation Gnatt Chart Video Recording	Worked on implementing Kmeans method Implemented a quantitative metric	Worked on new preprocessing methods as well as working on both models - Random Forest and GMM

Introduction:

The intersection of art and technology has garnered significant attention in recent years, mainly machine learning (ML) algorithms in creative domains. Personalized art recommendation systems enhance the user experience by culminating in the most suitable artworks for that particular individual. These systems utilize different methods to analyze user preferences and recommend artworks that resonate with the user. We will develop PaletteMatch—a model designed to analyze a user's art preferences and recommend pieces that align with their taste

This model will be predicated off of a WikiArt dataset containing thousands of paintings that has been labeled by Artist, Genre, and Style.

Dataset

Potential Results

Methods

References

Literature Review:

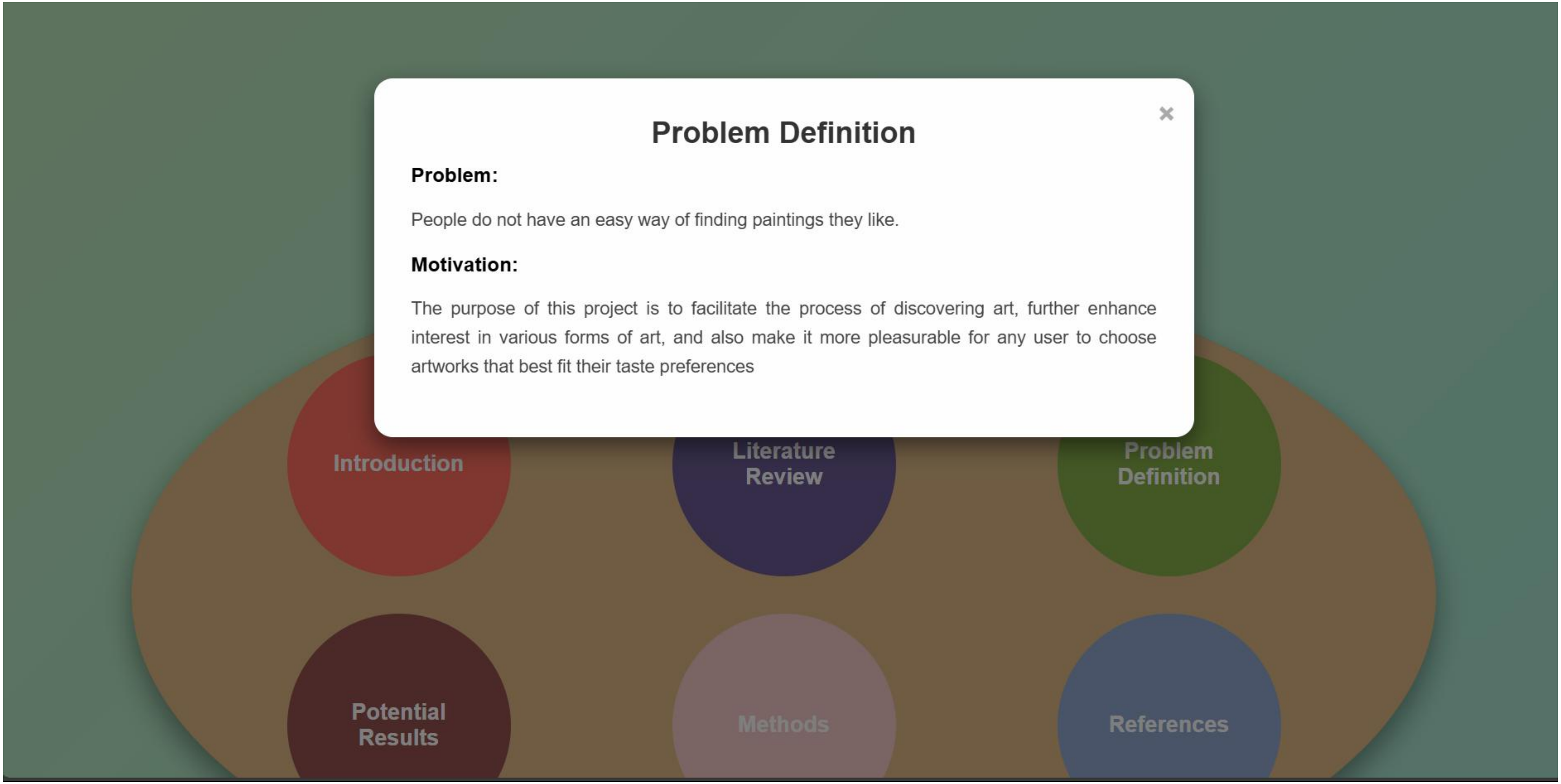
Roy Deepjyoti, Dutta Mala - A systematic review and research perspective on recommender systems. Journal of Big Data, 9(1), 97. This article discusses some of the issues about the evolution of automated recommendation systems, including data sparsity, bias, and deep learning trends, along with explainability. [3]

Messina, P., Dominguez, V., Parra, D., Trattner, C., & Soto, A. (2019). Content-based artwork recommendation: Integrating painting metadata with neural and manually-engineered visual features. The paper investigates how the metadata of paintings can be combined with neural and manually engineered visual features to enhance the performance of content-based artwork recommendation systems. We can use similar methodologies as mentioned in this study to advance the accuracy of personalized art suggestions. [2]

Potential Results

Methods

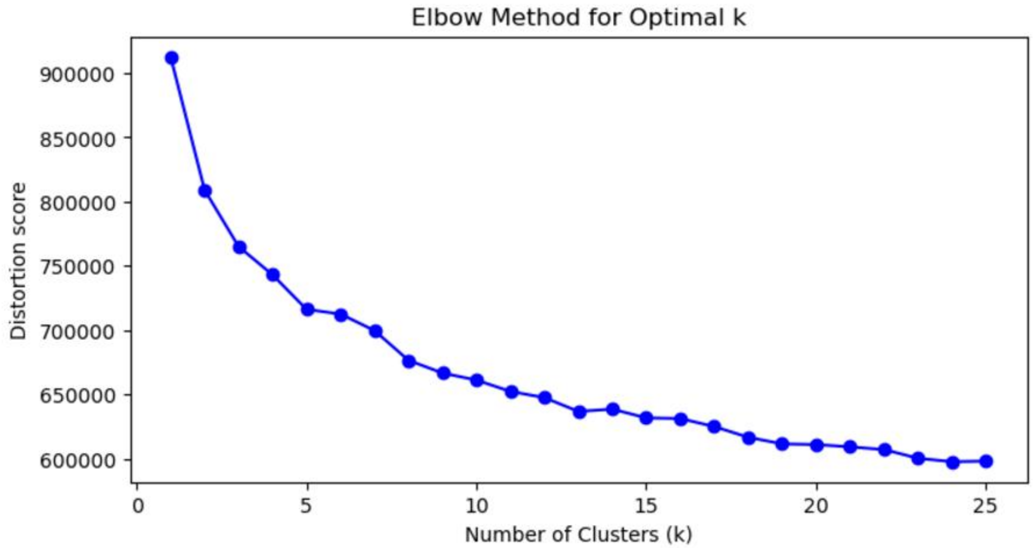
References


Problem Definition
Problem:
People do not have an easy way of finding paintings they like.
Motivation:
The purpose of this project is to facilitate the process of discovering art, further enhance interest in various forms of art, and also make it more pleasurable for any user to choose artworks that best fit their taste preferences
Introduction
Literature Review
Problem Definition
Potential Results
Methods
References

Results and Discussion

In this section, we analyze the clustering performance using various metrics and visualization methods.

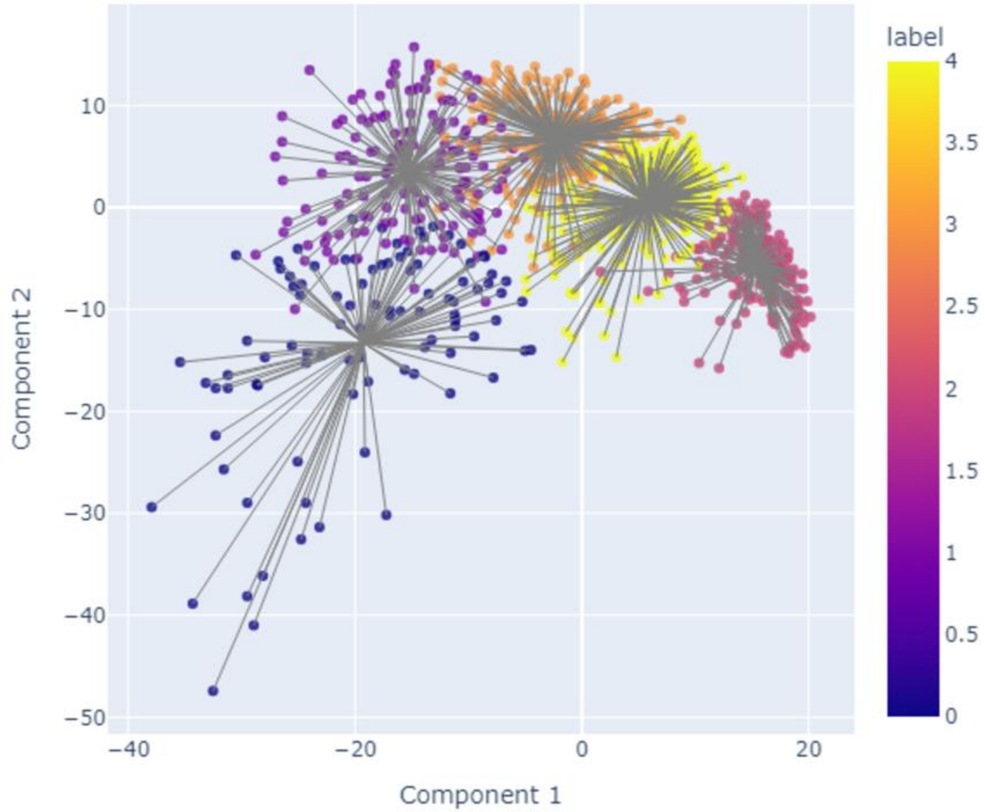
Elbow Method



We first did the elbow method to determine a good number of clusters we can seperate our images into. The elbow in our plot suggests that around 5 clusters were ideal as increasing more gives us diminishing returns.

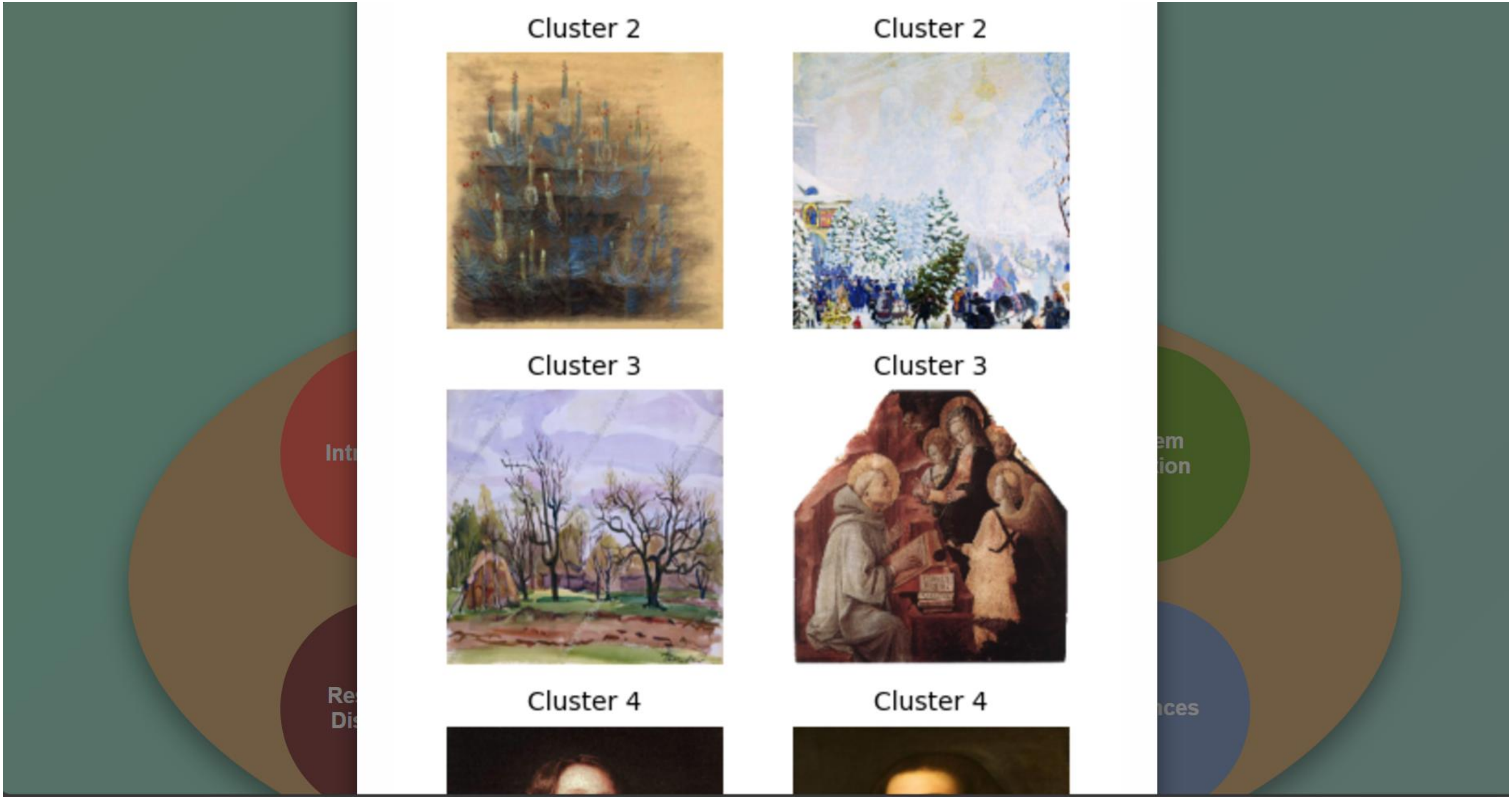
PCA Visualization of Clusters

Clusters Visualized Using PCA (Top 2 Components)



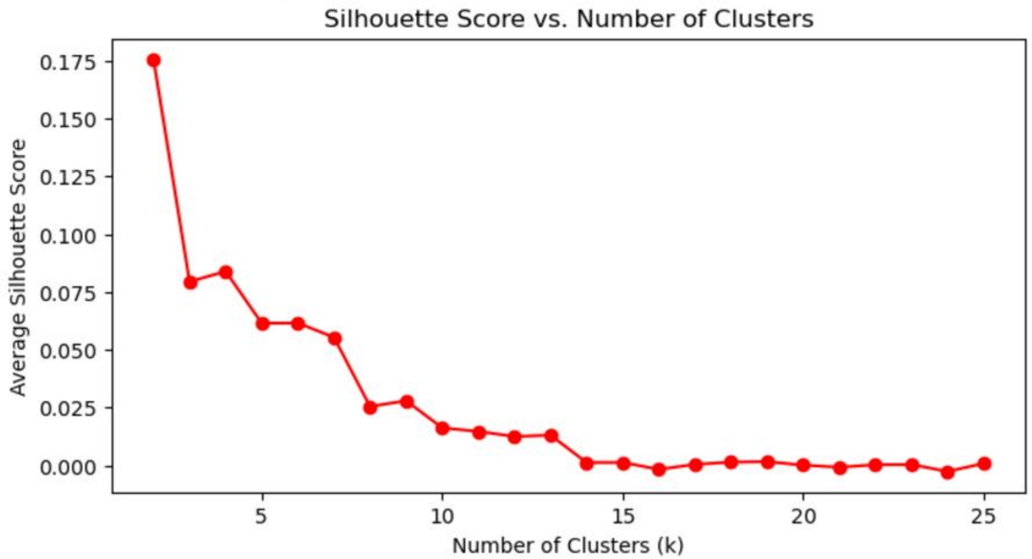
We ran kmeans on a subset of 1000 images from our dataset using the CNN extracted features of the 1000 images. We also used k-means++ to initialize our centroids better. And we used PCA for dimensionality reduction to help visualize the clusters in 2D space. The plot shows us that it indeed is possible to cluster the artworks and our kmeanss seperated clusters by their features to try and group similar artworks. However, there doesn't seem to be that much intercluster distance. This means that our features are not good enough at extracting differences between artworks. We plan on experimenting with other CNN's to extract features as well as using PCA before running Kmeans to reduce unimportant features.

Sample Images from Clusters


Cluster 2
Cluster 2
Cluster 3
Cluster 3
Cluster 4
Cluster 4

To see if the clusters were grouping similar images we plotted two images from each cluster to see if they resembled each other. It does seem that the clusters demonstrate a grouping of similar artistic styles. However sometimes the paintings in the same cluster don't resemble each other that well which again is probably because our feature representation of images was not good enough to make distinct clusters between images.

Silhouette Score Analysis



Methods

Data Preprocessing

Image Resizing and Normalization: Resize images to a fixed size and normalize pixel values. We chose this preprocessing method for two reasons. The first being that resizing ensures all images in our dataset will have the same dimensions. Neural networks require a fixed dimension per input and so we achieve that goal with this preprocessing method. The second reason is normalization, this helps to scale the pixel values which enhances the data come time to train the model. Essentially it ensures that higher pixel values do not unduly distort the model, just because they are bigger which makes the training process more optimized and generalizable. This is effectively avoiding any bias in the learning process.

Feature Extraction: Feature extraction includes the identification of important visual attributes from the images to improve clustering accuracy. In this project, we apply a Convolutional Neural Network to extract meaningful features like edges, textures, and patterns. CNNs are apt for this task since they automatically learn and represent visual hierarchies from simple to complex features. These extracted features are then used by clustering models, such as K-means and DBSCAN, to effectively group similar images for recommendations that are both accurate and visually relevant.

Models

K-means Clustering: To cluster paintings without labels. Note that we are using an unsupervised algorithm. We chose K-means because it groups the artwork based on their visual features. We are separating images into different clusters such that these groups of images have similar features, which are defined by PCA. Essentially the idea is that when a user inputs an images, K-means will enable the system to identify which cluster of images that image is most similar to, and rapidly recommend images from that cluster. This will help us to

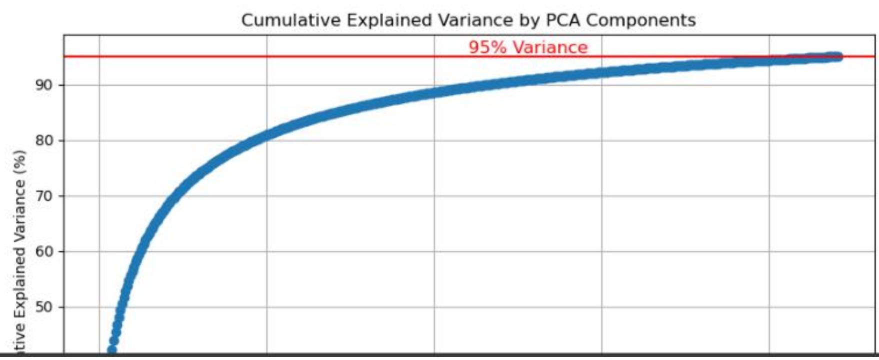
We used the Silhouette Score to help evaluate the separation between clusters and the cohesion within the cluster. A higher score indicates better-defined clusters. We plotted the silloute score over the number of clusters and we noticed as we increased clusters our silloute score decreased. At 5 clusters our silhouette score is quite low at 0.0185. This makes sense since our seperation between clusters is not good and we are going to try and improve this seperation in our final project.

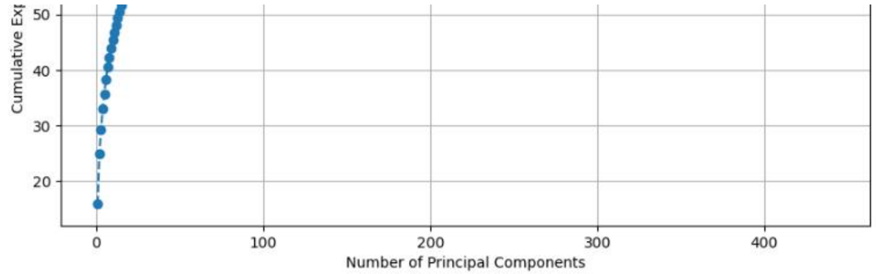
Davies-Bouldin Index

We also performed the Davies-Bouldin Index to also measure the cohesion and separation of the clusters. It gave us a score of 2.9782. This tells us that the clustering is not optimal as we are looking for a much smaller score.

Data Expansion and PCA Reduction

Before running our next models, we expanded our dataset to utilize all 80,000 images from the WikiArt dataset by processing and training using the PACE ICE supercomputing cluster. To improve feature representation, we reduced the number of features using PCA. To determine an optimal number of PCA components, we ran a script identifying the number of components retaining 95% variance.

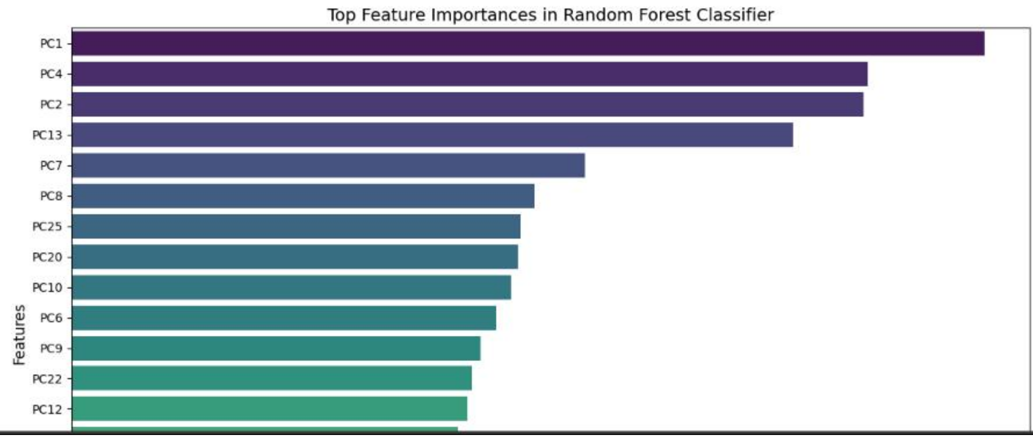


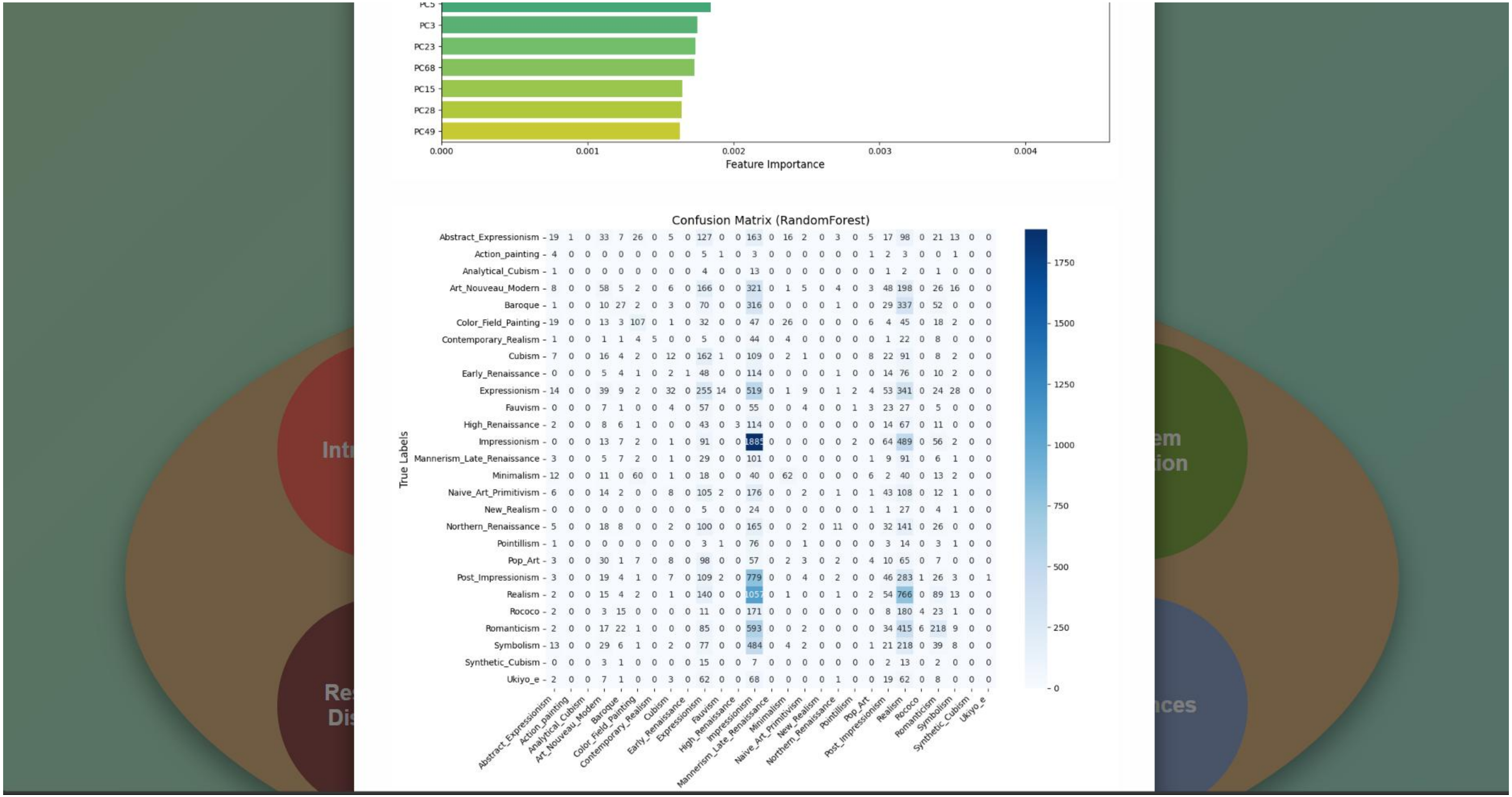


Random Forest

The Random Forest model achieved an accuracy of 21.4%, marginally better than random guessing. It effectively classified samples with decreasing Gini values down the tree, with leaf nodes achieving a Gini value of 0, indicating perfect classification for those samples.

Random Forest provides insights into feature importance. We plotted the top 10 most important features based on their Gini importance scores. The results indicated that PC1, PC4, PC2, and PC13 were the most significant, while other features showed lower importance, suggesting potential redundancy.




PC3
PC23
PC68
PC15
PC28
PC49
0.000
0.001
0.002
0.003
0.004
Feature Importance
Confusion Matrix (RandomForest)
True Labels
Abstract_Expressionism
Action_painting
Analytical_Cubism
Art_Nouveau_Modern
Baroque
Color_Field_Painting
Contemporary_Realism
Cubism
Early_Renaissance
Expressionism
Fauvism
High_Renaissance
Impressionism
Mannerism_Late_Renaissance
Minimalism
Naive_Art_Primitivism
New_Realism
Northern_Renaissance
Pointillism
Pop_Art
Post_Impressionism
Realism
Rococo
Romanticism
Symbolism
Synthetic_Cubism
Ukiyo_e
1750
1500
1250
1000
750
500
250
0

Diagonal cells in the confusion matrix represent correct predictions; other cells represent misclassifications. Styles like "Post-Impressionism" and "Romanticism" were classified more effectively due to noticeable contrasts.Significant confusion was observed between similar styles like "Abstract Expressionism" and "Expressionism."

Art Style	Precision	Recall	F1-Score	Support
Abstract Expressionism	0.15	0.03	0.06	556
Action Painting	0	0	0	20
Analytical Cubism	0	0	0	22
Art Nouveau Modern	0.16	0.07	0.09	867
Baroque	0.19	0.03	0.05	848
Color Field Painting	0.48	0.33	0.39	323
Contemporary Realism	1	0.05	0.1	96
Cubism	0.12	0.03	0.04	447
Early Renaissance	1	0	0.01	278
Expressionism	0.13	0.19	0.16	1347
Fauvism	0	0	0	187
High Renaissance	1	0.01	0.02	269
Impressionism	0.25	0.72	0.37	2612
Mannerism Late Renaissance	0	0	0	256
Minimalism	0.52	0.23	0.32	267
Naive Art Primitivism	0.05	0	0.01	481
New Realism	0	0	0	63
Northern Renaissance	0.39	0.02	0.04	510
Pointillism	0	0	0	103
Pop Art	0.09	0.01	0.02	297
Post Impressionism	0.08	0.04	0.05	1290
Realism	0.18	0.36	0.24	2147

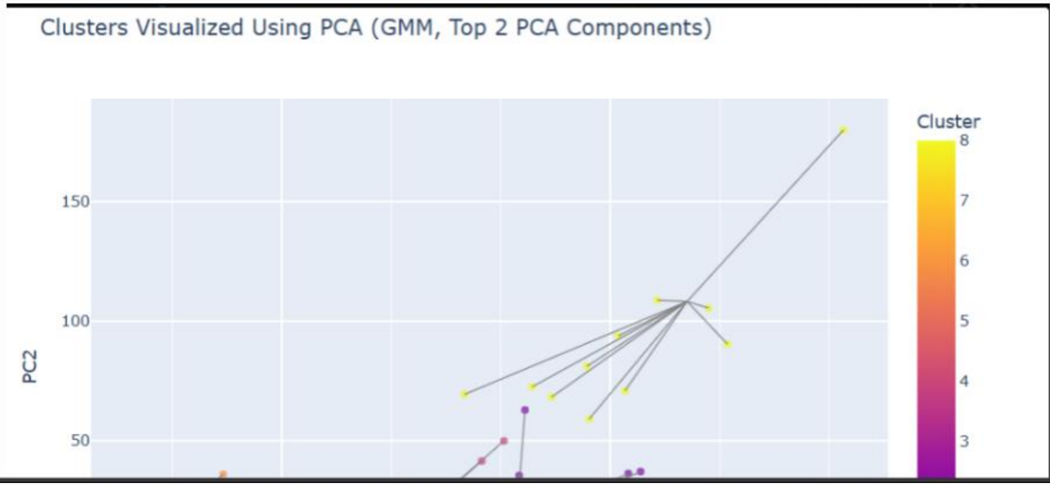
Realism	0.18	0.36	0.24	2147
Rococo	0.36	0.01	0.02	418
Romanticism	0.3	0.16	0.21	1404
Symbolism	0.08	0.01	0.02	905
Synthetic Cubism	0	0	0	43
Ukiyo-e	0	0	0	233

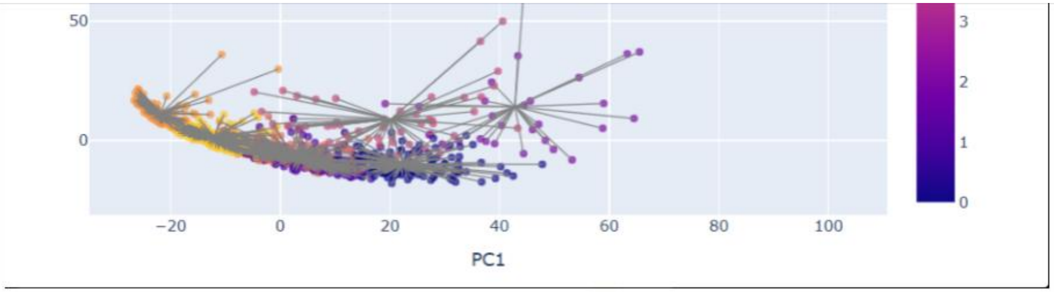
Several styles like Action Painting have zero performance metrics. High precision in few styles: Some styles like Contemporary Realism are predicted correctly but rarely.Impressionism is often identified but with many false positives. Some styles like Early Renaissance have low performance due to smaller sample sizes.

Overall Random Forest performed decent but not as we would have wanted.

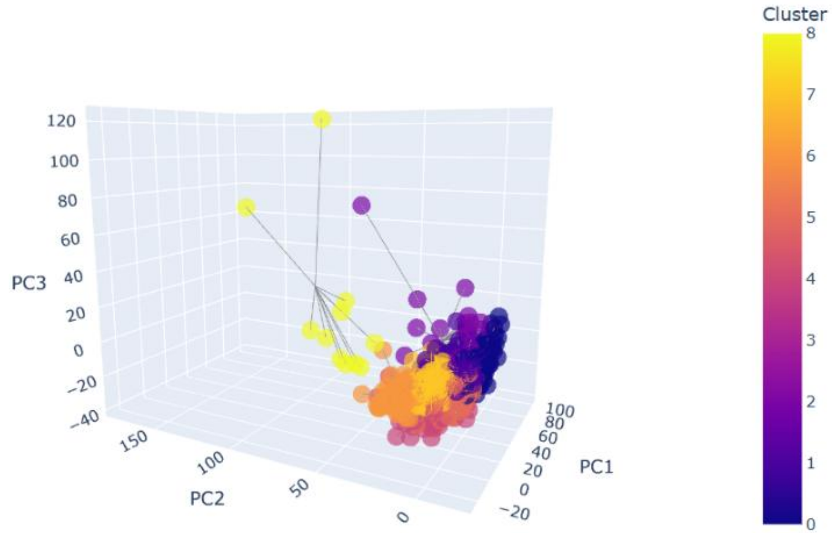
GMM

GMM was used to capture noncircular patterns in the data, but due to poor feature extraction and lack of labeled metadata, the results were suboptimal.

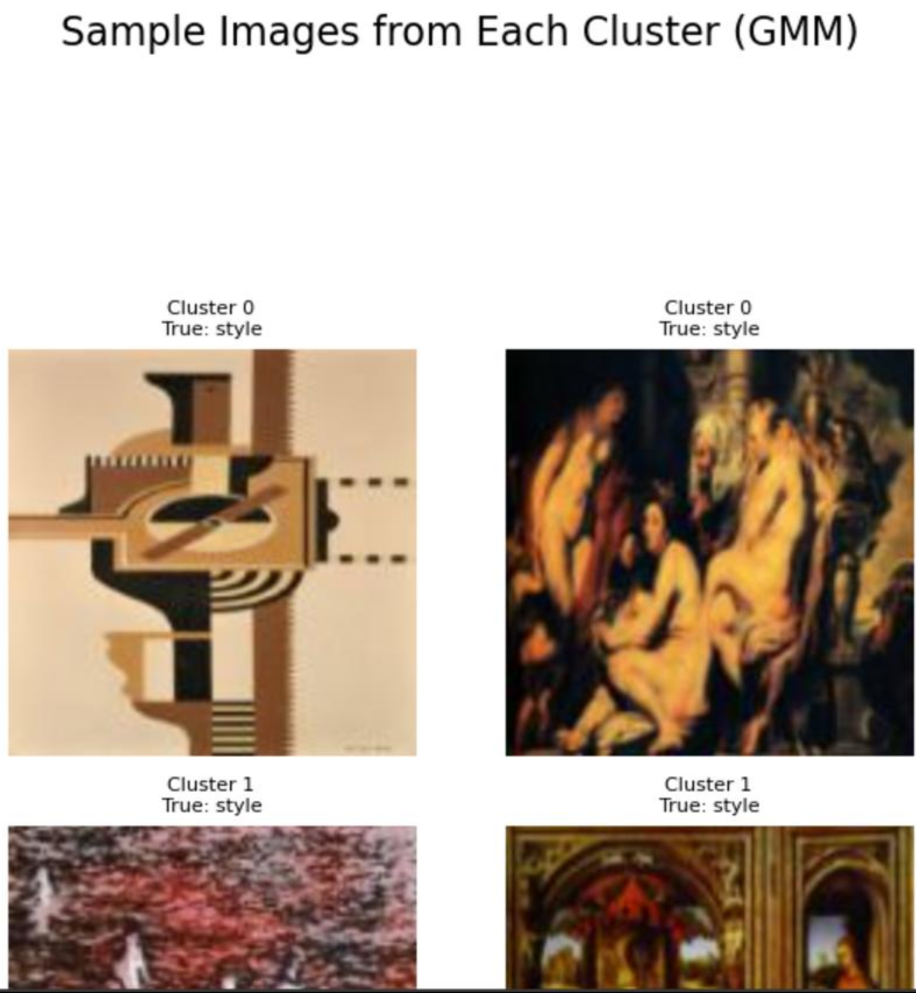


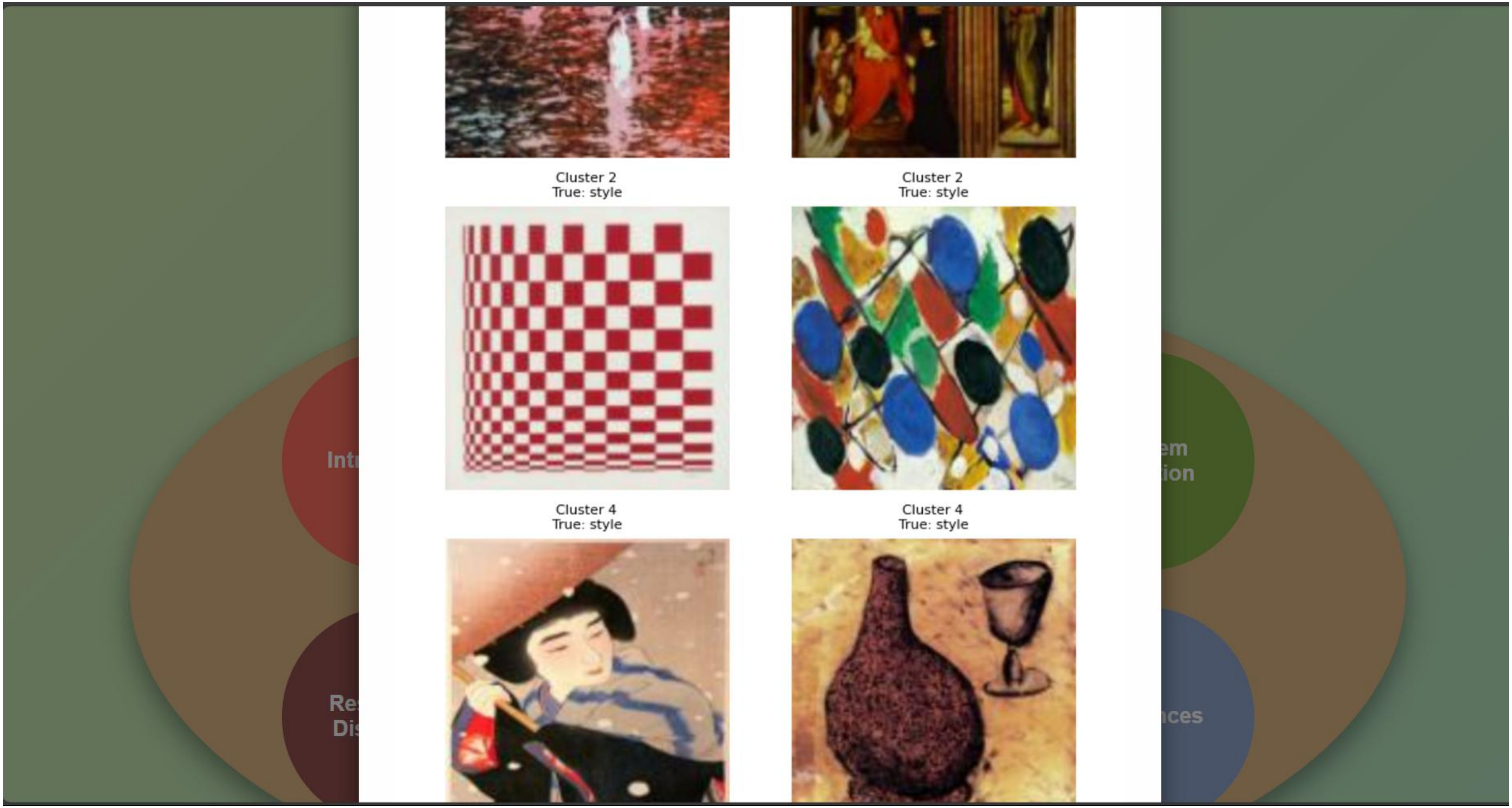


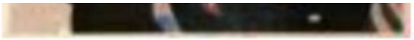
Clusters Visualized Using PCA (GMM, Top 3 PCA Components)



Higher-dimensional PCA revealed limited distinction between clusters, implying feature similarity across classes. Sample images from clusters also demonstrated inconsistent similarity across classes. Sample images from clusters also demonstrated inconsistent clustering due to insufficient feature representation.




Cluster 2
True: style
Cluster 2
True: style
Cluster 4
True: style
Cluster 4
True: style



Cluster 5
True: style

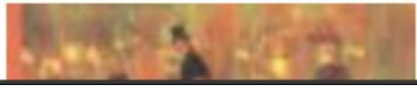
Cluster 5
True: style

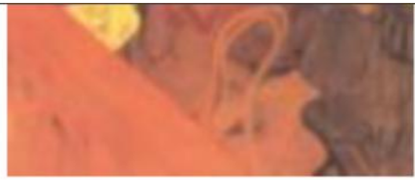
Cluster 6
True: style

Cluster 6
True: style

Cluster 7
True: style

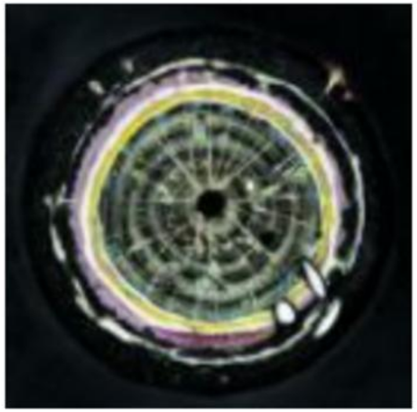
Cluster 7
True: style





Cluster 8
True: style

Cluster 8
True: style



Sample images from each cluster. Some clusters don't look too similar. Feature representation most likely isn't enough to make distinct clusters from images

Average Silhouette Score for KMeans (k=8): 0.0396 Silhouette Score for GMM (k=8): 0.0084 Davies-Bouldin Score for KMeans: 2.8245

As you can see there was not that much improvement with GMM based on the underlying reasons described above. So overall the quantitative metrics we utilized for this model showed very poor results in comparison to the other models, but improved when used with

Sample images from each cluster. Some clusters don't look too similar. Feature representation most likely isn't enough to make distinct clusters from images

Average Silhouette Score for KMeans (k=8): 0.0396 Silhouette Score for GMM (k=8): 0.0084 Davies-Bouldin Score for KMeans: 2.8245

As you can see there was not that much improvement with GMM based on the underlying reasons described above. So overall the quantitative metrics we utilized for this model showed very poor results in comparison to the other models, but improved when used with PCA. Art images generally follow complex, non-Gaussian distributions which leads to poor cluster definition and overlapping assignments over GMM. If we had better extracted features and metadata to use it would have been better to use GMM since there could be more distinctions.

Summary

Overall, while models like Random Forest, K-Means, and GMM show potential for classifying and recommending art, they face challenges like cluster definition, dimensionality reduction, and limitations in feature extraction. Future work involves using deeper CNNs, advanced clustering methods, and incorporating metadata like specific artist information to enhance results.

Models

<u>K-means Clustering</u>: To cluster paintings without labels. Note that we are using an unsupervised algorithm. We chose K-means because it groups the artwork based on their visual features. We are separating images into different clusters such that these groups of images have similar features, which are defined by PCA. Essentially the idea is that when a user inputs an images, K-means will enable the system to identify which cluster of images that image is most similar to, and rapidly recommend images from that cluster. This will help us to streamline the recommendation process.

<u>GMM</u>: The main clustering model is the Gaussian Mixture Model, which will be used to cluster paintings based on their visual features. GMM fits data in a probabilistic way, considering clusters as mixtures of Gaussian distributions, hence flexible and accurate modeling of complex datasets with clusters of arbitrary shape and size. GMM assigns each painting to a cluster by estimating the likelihood of belonging to different distributions using the previously determined optimal number of components. This ensures that GMM clusters paintings with nuanced and diverse features correctly, giving meaningful groupings that assist in downstream tasks such as image recommendations.

<u>Random Forest</u>: To classify paintings and predict labels from their visual features, we have chosen to use supervised learning algorithm, Random Forest. This model is particularly good in handling high-dimensional data and preventing overfitting. Specifically, in our application, images we use for our dataset typically have intricate visual patterns and diverse attributes that make classification challenging. Random This model will allow us to classify paintings into meaningful categories consisting of numerous style groups such as Abstract Impressionism.

Methods

Data Preprocessing

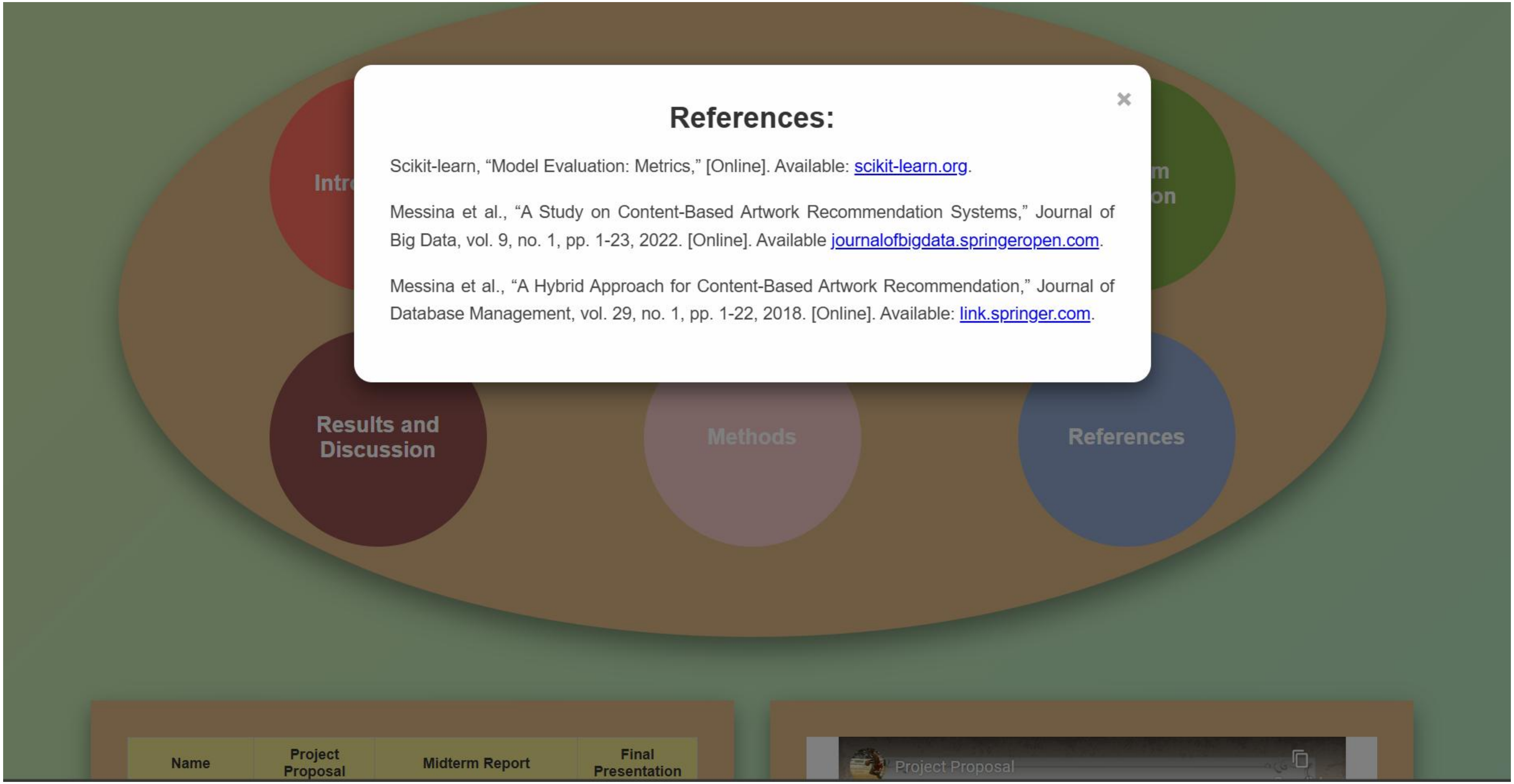
Image Resizing and Normalization: Resize images to a fixed size and normalize pixel values. We chose this preprocessing method for two reasons. The first being that resizing ensures all images in our dataset will have the same dimensions. Neural networks require a fixed dimension per input and so we achieve that goal with this preprocessing method. The second reason is normalization, this helps to scale the pixel values which enhances the data come time to train the model. Essentially it ensures that higher pixel values do not unduly distort the model, just because they are bigger which makes the training process more optimized and generalizable. This is effectively avoiding any bias in the learning process.

Models

K-means Clustering: To cluster paintings without labels. Note that we are using an unsupervised algorithm. We chose K-means because it groups the artwork based on their visual features. We are separating images into different clusters such that these groups of images have similar features, which are defined by PCA. Essentially the idea is that when a user inputs an images, K-means will enable the system to identify which cluster of images that image is most similar to, and rapidly recommend images from that cluster. This will help us to streamline the recommendation process.

Name	Project Proposal	Midterm Report	Final Presentation

Project Proposal


×
References:
Scikit-learn, “Model Evaluation: Metrics,” [Online]. Available: scikit-learn.org.
Messina et al., “A Study on Content-Based Artwork Recommendation Systems,” Journal of Big Data, vol. 9, no. 1, pp. 1-23, 2022. [Online]. Available journalofbigdata.springeropen.com.
Messina et al., “A Hybrid Approach for Content-Based Artwork Recommendation,” Journal of Database Management, vol. 29, no. 1, pp. 1-22, 2018. [Online]. Available: link.springer.com.
Results and Discussion
Methods
References
Name
Project Proposal
Midterm Report
Final Presentation
Project Proposal