DBSCAN & GMM Clusterization for Vegetable Dataset

Introduction

This study investigates unsupervised classification methods, specifically Density-Based Spatial Clustering of Applications with Noise (DBSCAN) and Gaussian Mixture Models (GMM), applied to a vegetable dataset. Various feature extraction techniques, including Histogram of Oriented Gradients (HOG), RGB Histogram, and custom-trained Autoencoders (VQ-VAE and VAE), are employed to extract informative features from the dataset. Additionally, supervised classification models such as Support Vector Machines (SVM) and Random Forest, along with a Dummy Classifier for random chance comparison, are utilized for benchmarking purposes. A comprehensive grid search is performed to explore a wide range of hyperparameter values for the clustering and classification models.

Methodology

Data selection

The dataset comprised 15 distinct vegetable classes, each containing 1000 training images and 200 images for both validation and test sets. However, for pragmatic reasons, only two classes were selected for analysis. This decision was motivated by two primary considerations.

Firstly, certain classes exhibited remarkable visual similarity, sharing analogous colour palettes and shapes, with subtle pattern differences that could be subjectively influenced by image clarity. To mitigate potential ambiguities in classification, two classes were chosen to enhance the clarity of distinctiveness between clusters.

Secondly, the choice to focus on two classes also addressed computational efficiency concerns during the hyperparameter tuning phase. The grid search process, particularly for models such as Gaussian Mixture Models (GMM) applied to the final selected features, required extensive computational resources. By reducing the dataset size to 2000 images (from the original 15,000), the grid search time was substantially decreased. For instance, the GMM search for the chosen features was completed in approximately 5 hours, enabling a more expedited exploration of hyperparameter configurations while maintaining a representative dataset size. This strategic reduction in classes and images facilitated a more manageable and efficient model optimization process.

Feature extraction

- **Histogram of Oriented Gradients (HOG):**

To expedite computation, images were resized to 64x64 pixels for HOG extraction. HOG was computed with 9 bins in cells of 8x8 pixels and 2x2 cells per block. Notably, HOG was computed for both the green and red channels, and the results were concatenated. While initial experiments used only the grayscale channel, combining red and green channels proved more effective, likely due to the distinctive colour characteristics of the dataset.

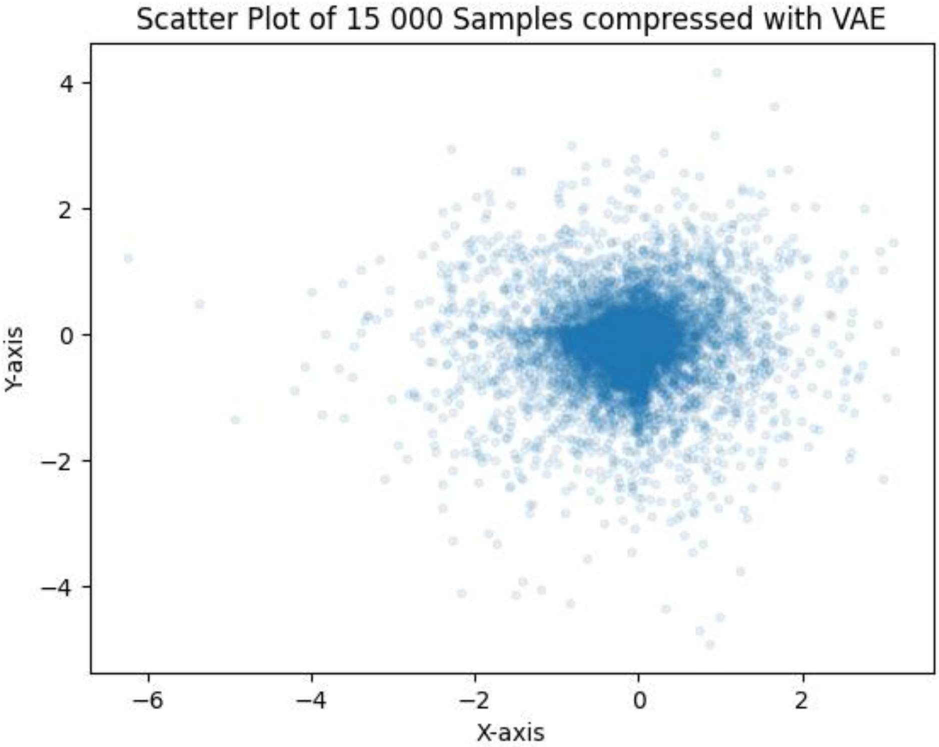
- **RGB Histogram:**

A simple RGB histogram was computed, capturing the distribution of pixel values in each channel (red, green, blue). Given the dominant colours in the dataset (green and orange), the expectation was to observe Gaussian-like distributions in the histogram. Reducing computational load and aligning with dataset characteristics, only the red and green channels were considered for analysis.

- **Vector Quantized Variational Autoencoder (VQ-VAE):**

Implementing the architecture outlined in the "Generating Diverse High-Fidelity Images with VQ-VAE-2" paper, VQ-VAE featured a top quantized block of 28x28x1 and a bottom quantized block of 14x14x1, with a learnable codebook of 1024 values. A mean squared error (MSE) loss of 0.02 was achieved for the reconstruction task. While experiments with a larger top block yielded lower MSE values, the subsequent clusterization results favoured the more compressed bottom block.

A Variational Autoencoder (VAE) was also employed, compressing the 14x14x1 features to a 2-dimensional bottleneck, capturing mean and variance. However, the KL-divergence constraint in the VAE led to a well-mixed 2D Gaussian distribution, inhibiting distinct class clusterization, as I initially expected.



These diverse feature extraction techniques aimed to capture different aspects of the vegetable dataset, considering colour, shape, and latent representations, and formed the basis for subsequent unsupervised and supervised classification analyses.

Supervised Classification

- **Support Vector Machine & Random Forest:**

For supervised classification, Support Vector Machine (SVM) and Random Forest models were employed on all types of features. The SVM models were fine-tuned by exploring various C values (0.001, 0.01, 0.1, 1, 10, 100, 1000), and Random Forest models were fine-tuned by varying the number of estimators (50, 100, 200, 300). The fine-tuning process focused on maximizing validation accuracy.

The best-performing models, determined based on validation accuracy, were then evaluated on the test dataset to assess their generalization performance. The metrics obtained from these evaluations served as the final performance indicators for each feature extraction method.

Metrics:

- HOG:
  - SVM [C=10]: 0.98% accuracy
  - RF [100 estimators]: 92% accuracy
- RGB Histogram:

- SVM [C=1]: 99% accuracy
- RF [50 estimators]: 99% accuracy

- VQ-VAE:
  - SVM [C=10]: 99% accuracy
  - RF [300 estimators]: 99% accuracy
- VAE:
  - SVM [C=0.001]: 50% accuracy *(*on validation)*
  - RF [100 estimators]: 51% accuracy *(*on validation)*

** Note on VAE Features:*

In the case of features extracted using the Variational Autoencoder (VAE), test metrics were not performed. The validation results already indicated that this feature extraction approach tended to mix classes into the same distribution, making it challenging for supervised classification methods to differentiate between classes. Therefore, the VAE features were excluded from the final performance assessment, acknowledging their limitations in class separability. The focus on validation results sufficed to justify the impracticality of using VAE features for supervised classification in this context.

Random Chance

To establish a baseline for comparison, a Dummy Classifier from the sklearn library was employed with various strategies. The obtained scores for each strategy are as follows:

Strategy: most_frequent	Accuracy: 50%
Strategy: prior	Accuracy: 50%
Strategy: stratified	Accuracy: 47.75%
Strategy: uniform	Accuracy: 49.75%

The results provide a reference point for evaluating the performance of the unsupervised and supervised classification models, allowing for a meaningful comparison against random chance in subsequent analyses.

DBSCAN

For DBSCAN (Density-Based Spatial Clustering of Applications with Noise), the critical parameters are eps and min_samples. These parameters play crucial roles in defining clusters within the data:

- eps (Epsilon): It represents the maximum distance between two samples for one to be considered as in the neighbourhood of the other. In other words, it defines the radius within which the algorithm looks for neighbouring points.
- min_samples: It specifies the minimum number of samples required to form a dense region, i.e., the minimum number of data points within the specified eps distance to consider a point as a core point.

Grid Search for Optimal Parameters:

To determine a suitable range for eps in the grid search, the Euclidean distances between the training data points and their shuffled counterparts were calculated. Multiple iterations of this process were performed to create plots, resulting in a narrowed range of possible eps values that demonstrated promising clusterization potential.

For min_samples, given the dataset's size of 2000 samples, a reasonable range was selected, varying from a minimum of 5 to a maximum of 500.

Silhouette Score and Constraints:

The Silhouette Score was employed as the metric for evaluating cluster quality. However, in high-dimensional spaces, DBSCAN tends to favour either numerous small clusters or a single large cluster, affecting the Silhouette Score. To address this, two constraints were implemented:

- Cluster Size Constraint: Only cauterizations with a maximum of 8 clusters were considered.
- Dataset Fraction Constraint: The maximum cluster size was limited to a fraction of the whole dataset. Individual searches were conducted for fractions of 0.55, 0.6, 0.65, 0.7, and 0.75.

These constraints aimed to mitigate the impact of high dimensionality and foster more meaningful and interpretable cauterizations. The resulting DBSCAN configurations were then compared based on the Silhouette Score, offering insights into the algorithm's performance under different parameter settings.

Integration with Supervised Classification and Label Propagation:

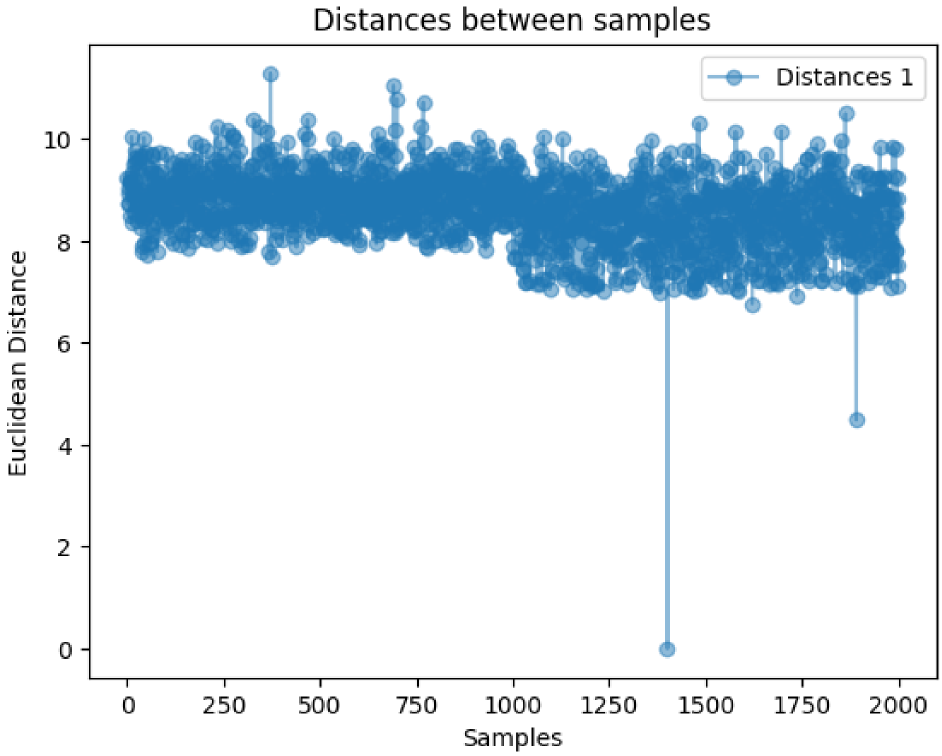
In the final phase, the DBSCAN clusterization results were compared with supervised classification outcomes on the test set. However, a challenge emerged as DBSCAN lacks an individual predict function for test data.

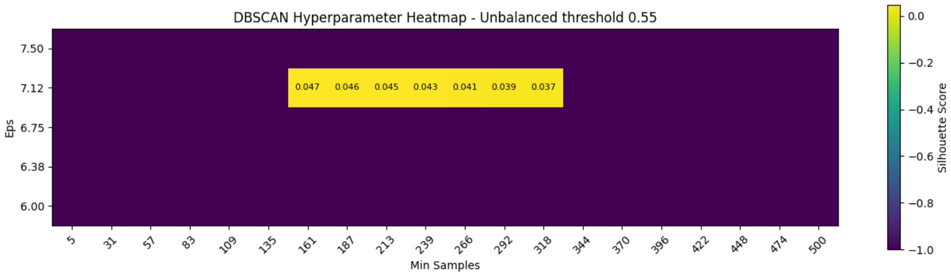
To bridge this gap, a simple approach was devised: for each test sample, the closest sample from the training set was identified, and the corresponding training cluster label was assigned to the test sample. This process essentially mirrors a k-nearest neighbours (KNN) approach with k set to 1.

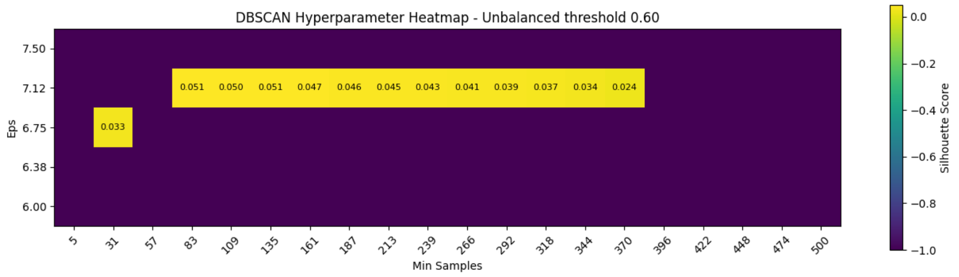
Recognizing the potential impact of the choice of k on the overall classification performance, an extension was explored. A fine-tuning process was conducted to optimize the k value, further refining the label propagation strategy. This extension aimed to enhance the classification accuracy and align the performance of DBSCAN with that of the supervised classification models, providing a comprehensive understanding of the clustering and classification dynamics on the vegetable dataset.

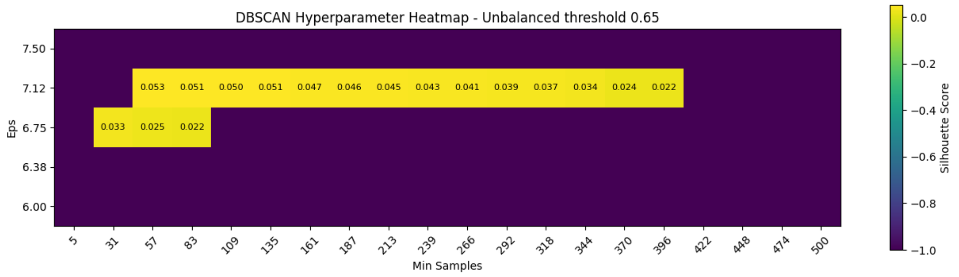
Furthermore, to facilitate a direct comparison between ground truth labels and cluster labels under the specified constraints (maximum 8 clusters and varying dataset fractions), a straightforward permutation method was employed. Given the consistency in obtaining only two clusters within the defined constraints, this permutation technique ensured a seamless alignment between ground truth labels and cluster labels, enabling a meaningful evaluation of the classification results.

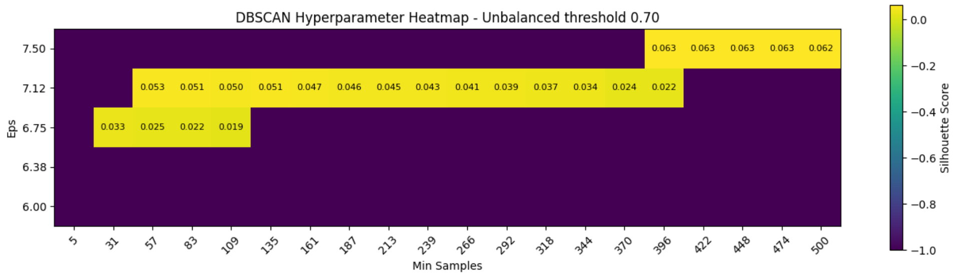
- **HOG**

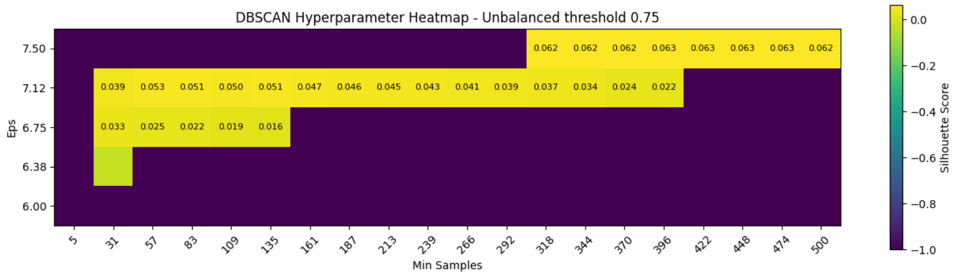


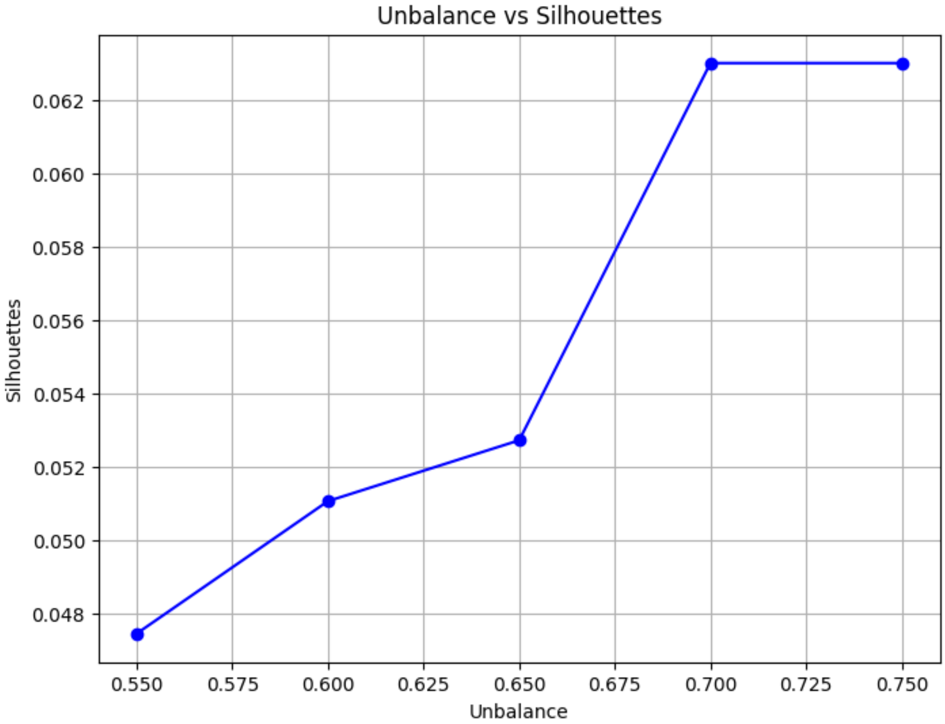


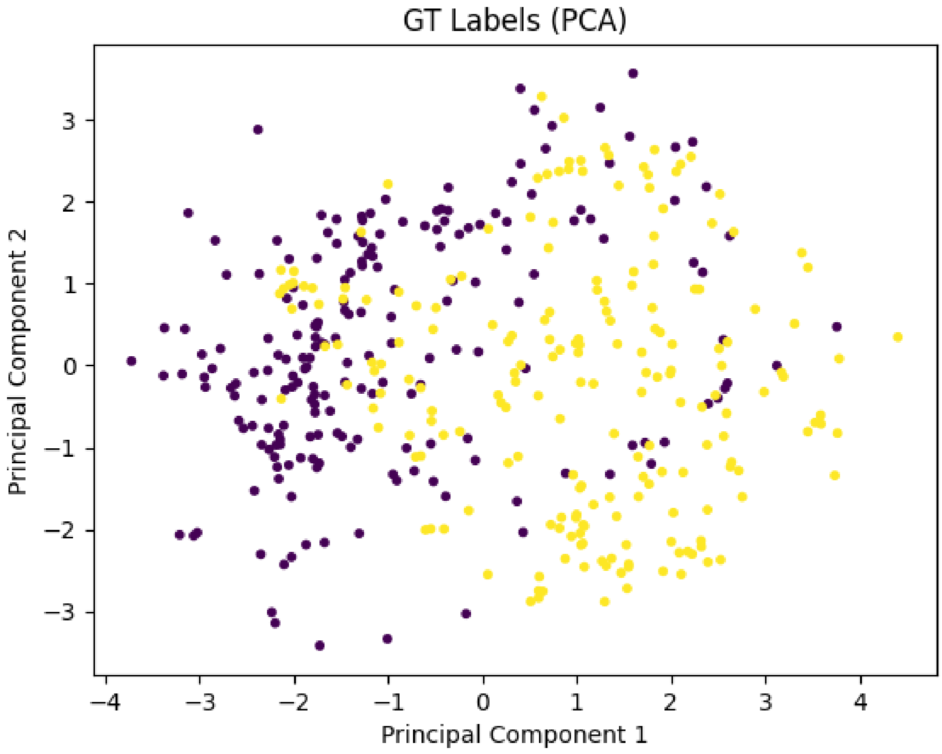

DBSCAN Hyperparameter Heatmap - Unbalanced threshold 0.60
Eps
7.50
7.12
6.75
6.38
6.00
0.051 0.050 0.051 0.047 0.046 0.045 0.043 0.041 0.039 0.037 0.034 0.024
0.033
5 31 57 83 109 135 161 187 213 239 266 292 318 344 370 396 422 448 474 500
Min Samples
Silhouette Score
0.0 −0.2 −0.4 −0.6 −0.8 −1.0


DBSCAN Hyperparameter Heatmap - Unbalanced threshold 0.65
Eps
7.50
7.12
6.75
6.38
6.00
0.053 0.051 0.050 0.051 0.047 0.046 0.045 0.043 0.041 0.039 0.037 0.034 0.024 0.022
0.033 0.025 0.022
5 31 57 83 109 135 161 187 213 239 266 292 318 344 370 396 422 448 474 500
Min Samples
Silhouette Score
0.0 −0.2 −0.4 −0.6 −0.8 −1.0

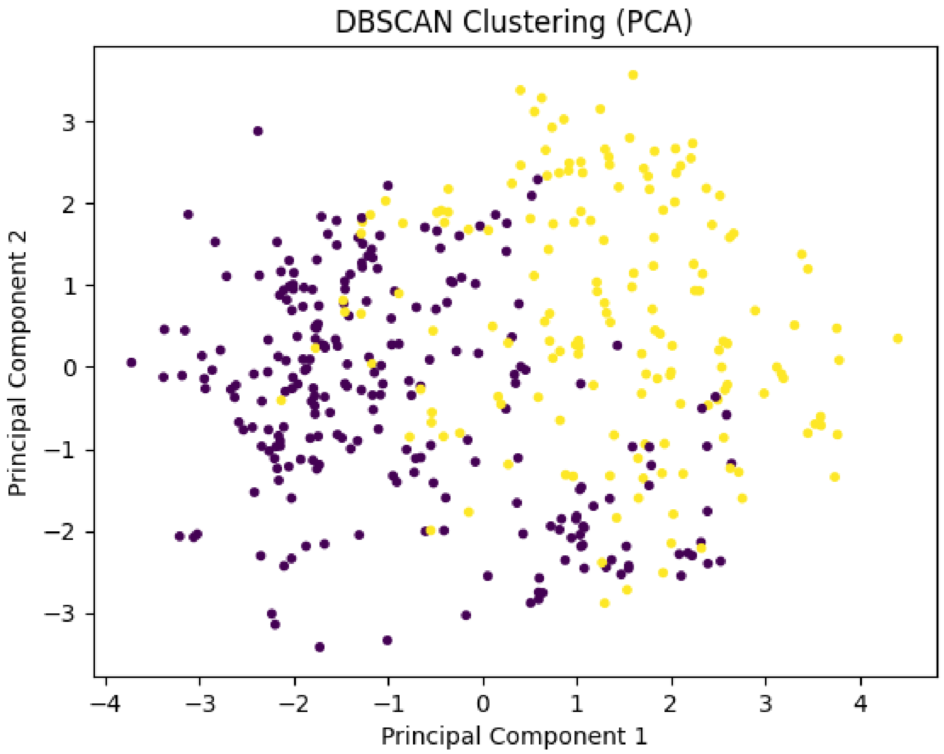

DBSCAN Hyperparameter Heatmap - Unbalanced threshold 0.70
Eps
7.50
7.12
6.75
6.38
6.00
0.063 0.063 0.063 0.063 0.062
0.053 0.051 0.050 0.051 0.047 0.046 0.045 0.043 0.041 0.039 0.037 0.034 0.024 0.022
0.033 0.025 0.022 0.019
5 31 57 83 109 135 161 187 213 239 266 292 318 344 370 396 422 448 474 500
Min Samples
Silhouette Score
0.0 −0.2 −0.4 −0.6 −0.8 −1.0

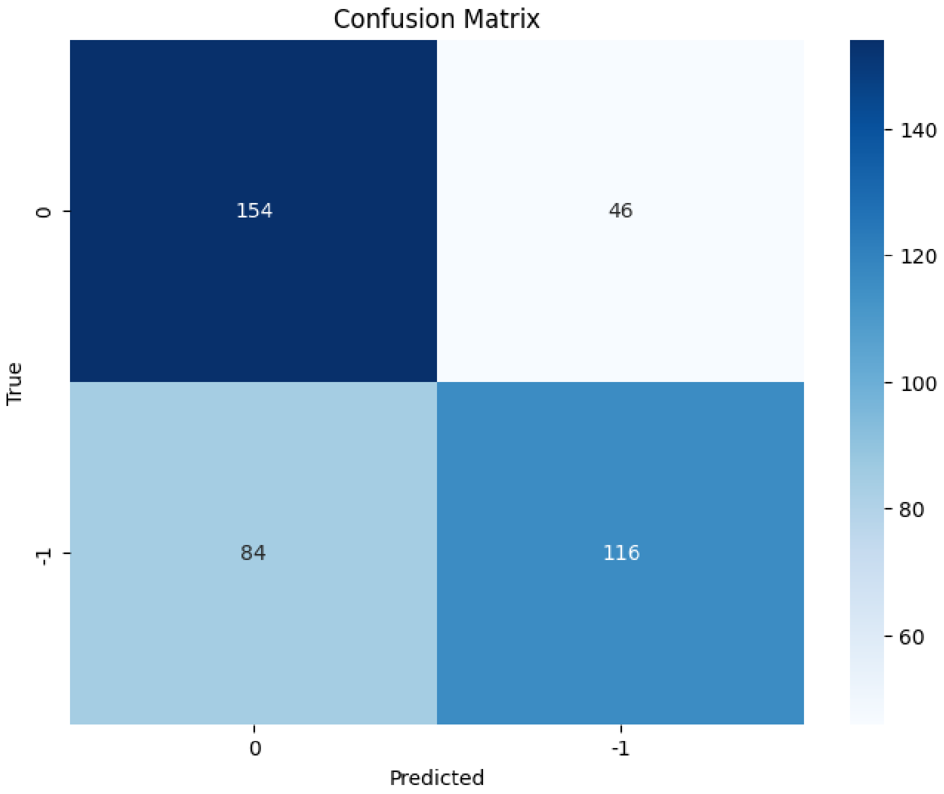

DBSCAN Hyperparameter Heatmap - Unbalanced threshold 0.75
Eps
7.50
7.12
6.75
6.38
6.00
0.062 0.062 0.062 0.063 0.063 0.063 0.063 0.062
0.039 0.053 0.051 0.050 0.051 0.047 0.046 0.045 0.043 0.041 0.039 0.037 0.034 0.024 0.022
0.033 0.025 0.022 0.019 0.016
5 31 57 83 109 135 161 187 213 239 266 292 318 344 370 396 422 448 474 500
Min Samples
Silhouette Score
0.0 −0.2 −0.4 −0.6 −0.8 −1.0



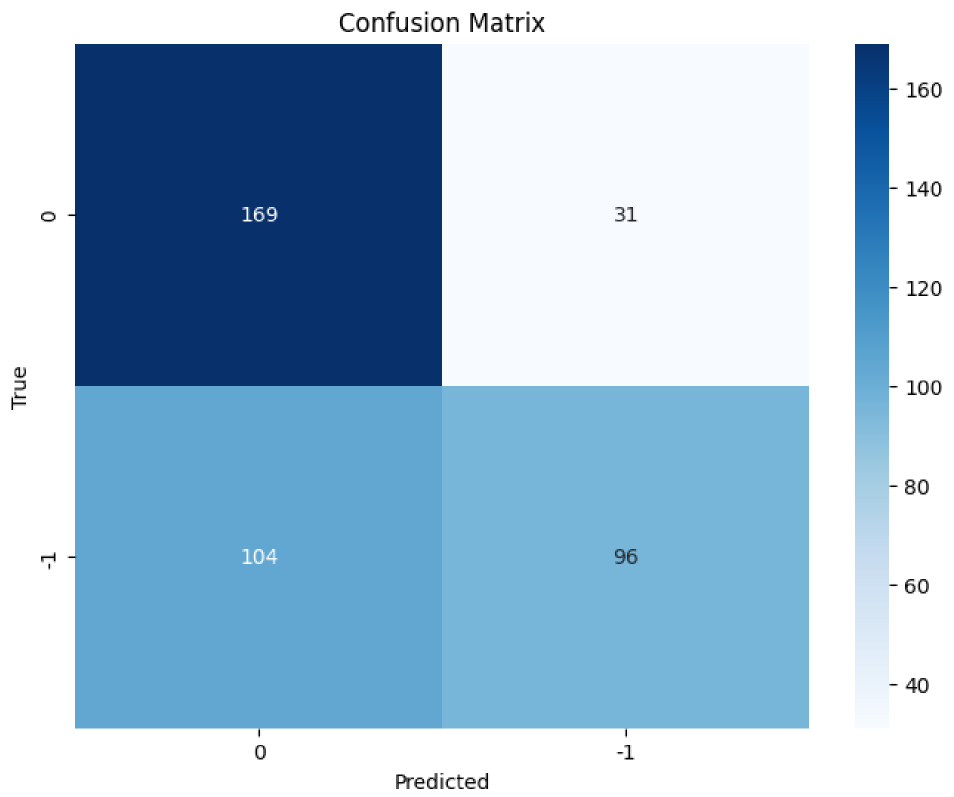


- Eps: 7.12 - Min: 161 - **Threshold 0.55**
  - Silhouette Score: **0.046**
  - Accuracy: **68%**

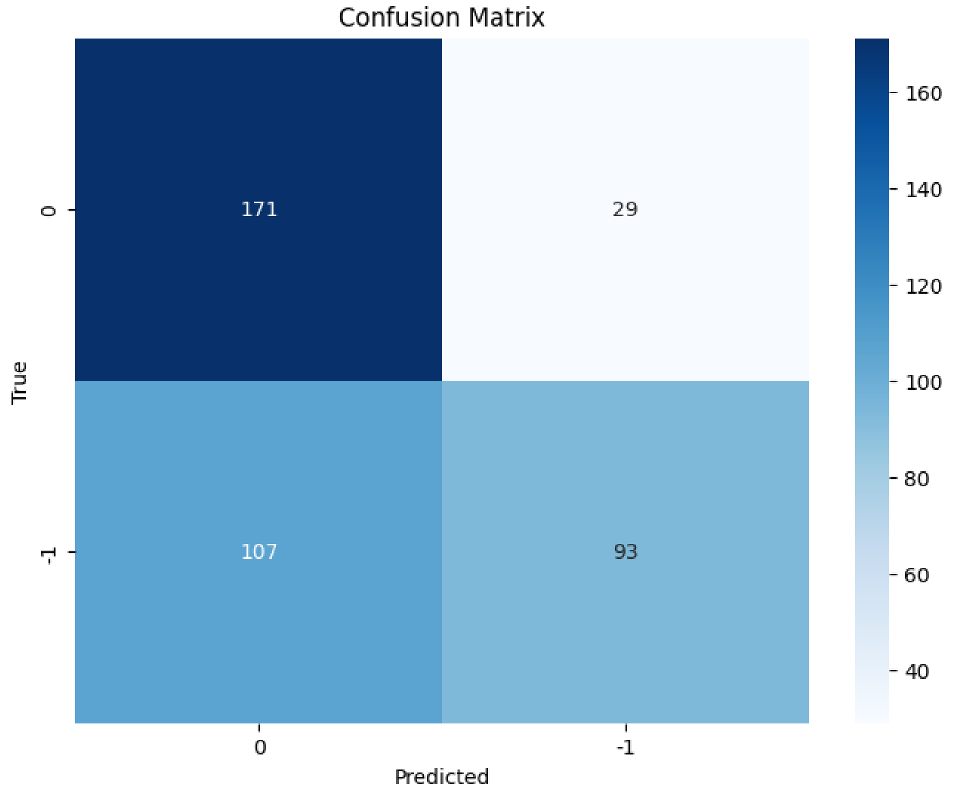




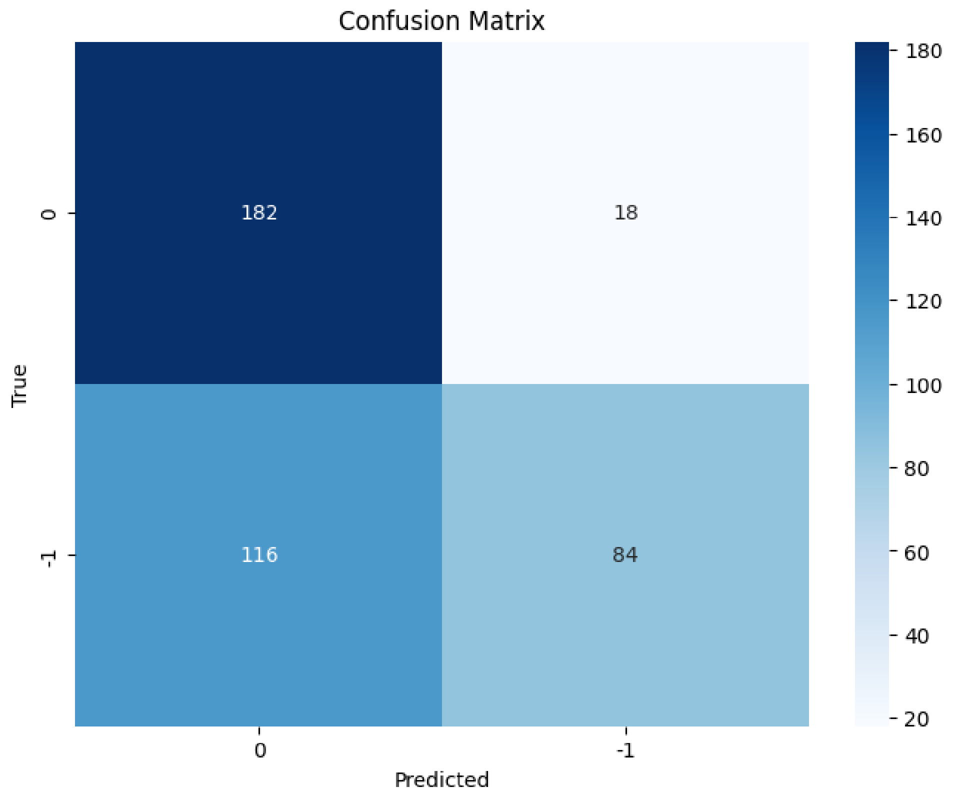
- Eps: 7.12 - Min: 83 - **Threshold 0.6**
  - Silhouette Score: **0.05**
  - Accuracy: **66%**



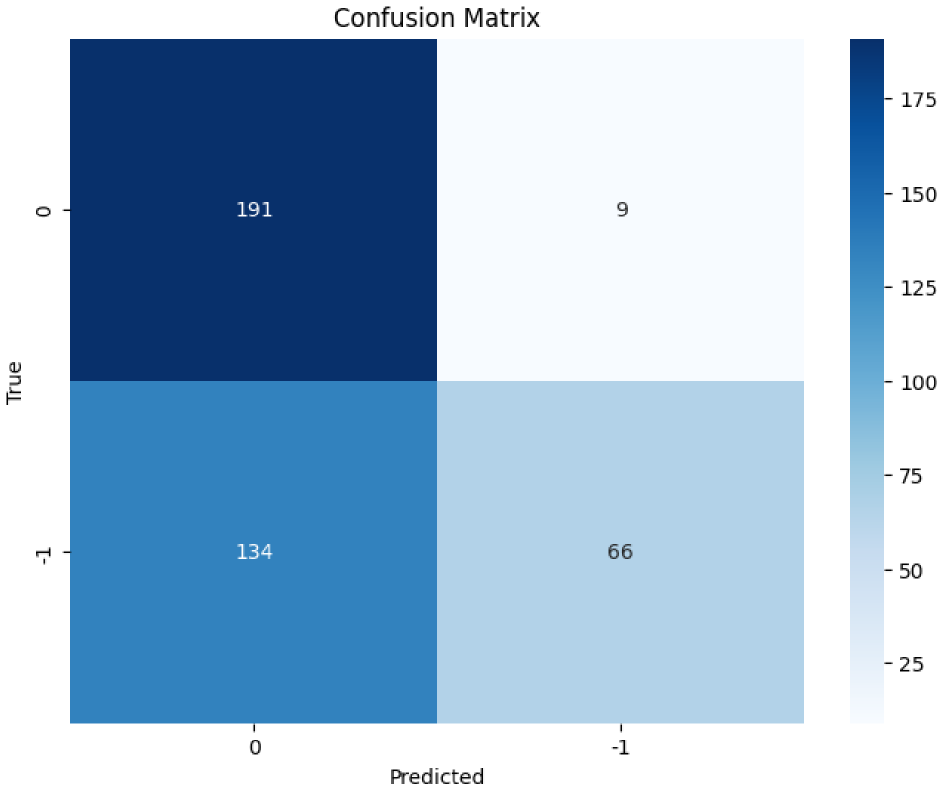
- Eps: 7.12 - Min: 57 - **Threshold 0.65**
    - Silhouette Score: **0.051**
    - Accuracy: **66%**



- Eps: 7.12 - Min: 396 - **Threshold 0.7**
  - Silhouette Score: **0.062**
  - Accuracy: **67%**

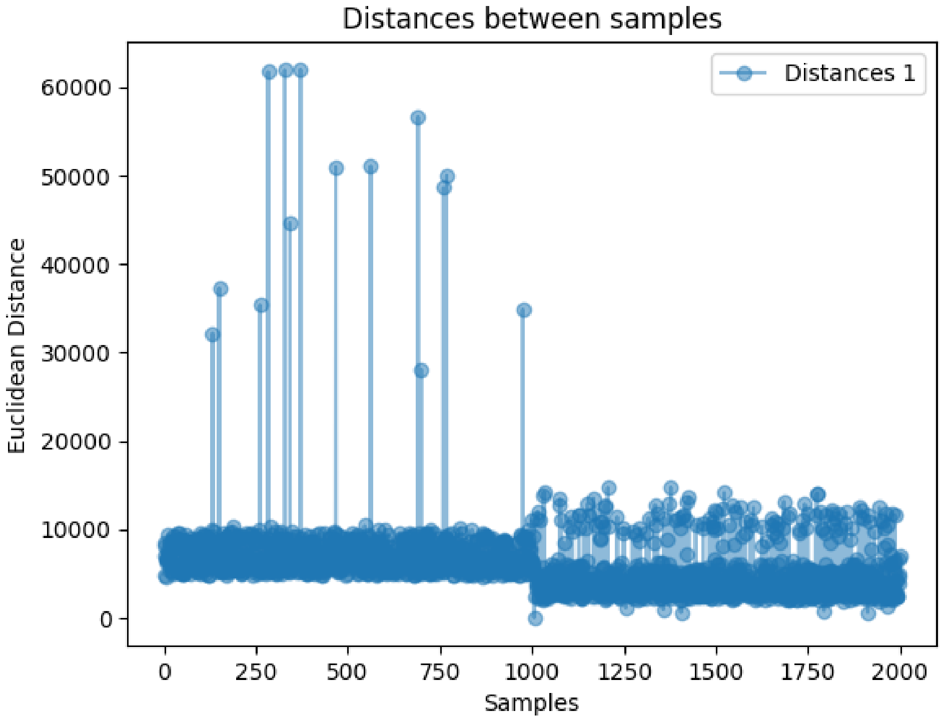


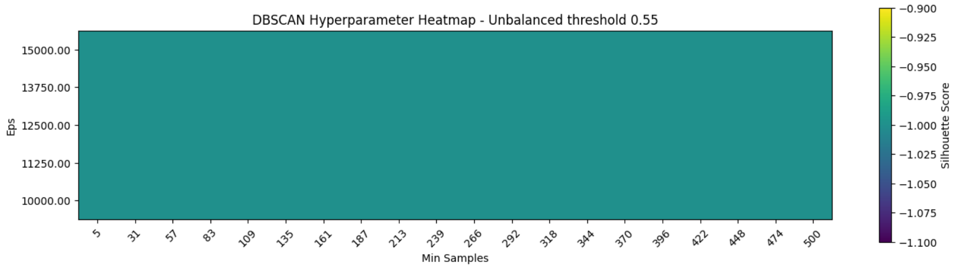
- Eps: 7.12 - Min: 306 - **Threshold 0.75**
  - Silhouette Score: **0.064**
  - Accuracy: **64%**

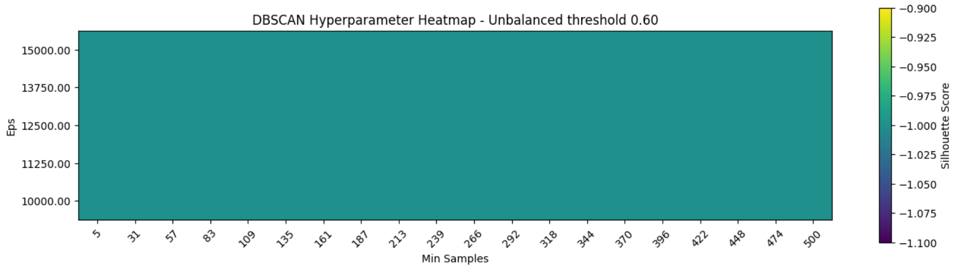


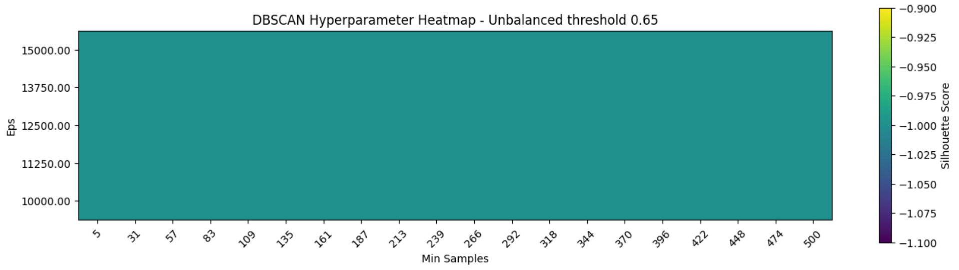
- **RGB Histogram**

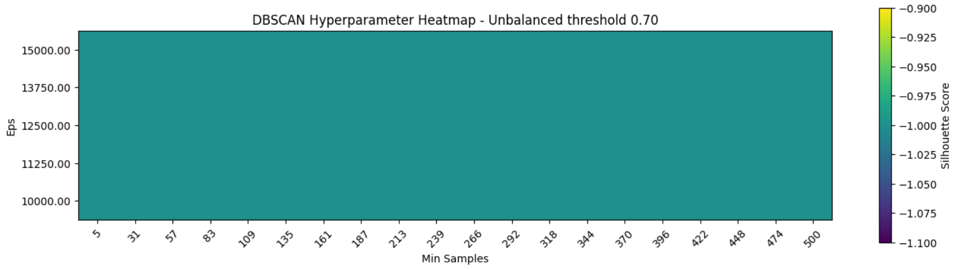
There is no model that gets a silhouette score higher than 0, under this grid search. As so, I did not further perform any accuracy performance for this type of features.

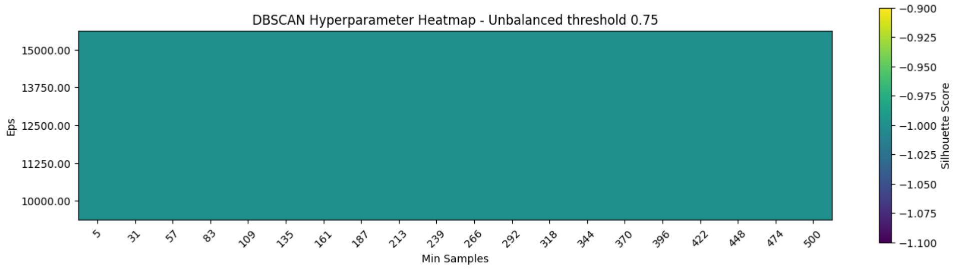



DBSCAN Hyperparameter Heatmap - Unbalanced threshold 0.55
15000.00
13750.00
12500.00
11250.00
10000.00
Eps
5 31 57 83 109 135 161 187 213 239 266 292 318 344 370 396 422 448 474 500
Min Samples
−0.900 −0.925 −0.950 −0.975 −1.000 −1.025 −1.050 −1.075 −1.100
Silhouette Score

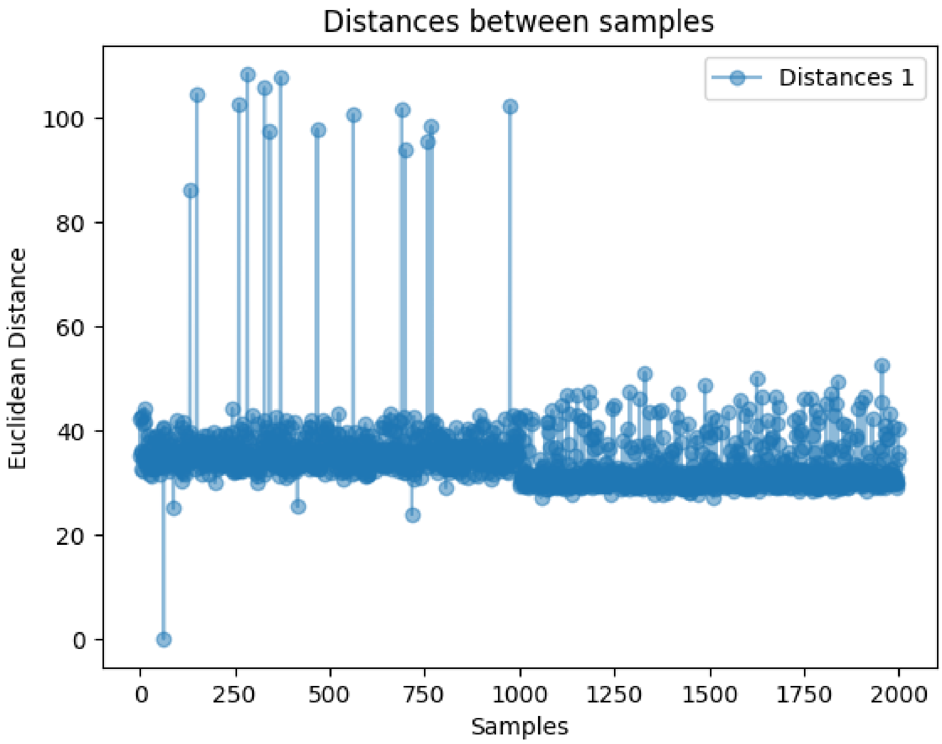

DBSCAN Hyperparameter Heatmap - Unbalanced threshold 0.60
15000.00
13750.00
12500.00
11250.00
10000.00
Eps
5 31 57 83 109 135 161 187 213 239 266 292 318 344 370 396 422 448 474 500
Min Samples
−0.900 −0.925 −0.950 −0.975 −1.000 −1.025 −1.050 −1.075 −1.100
Silhouette Score

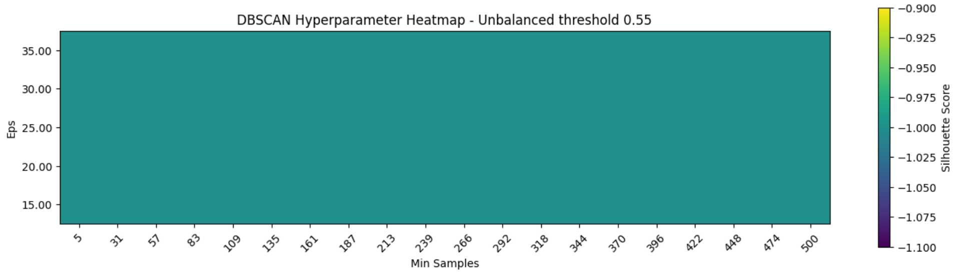

DBSCAN Hyperparameter Heatmap - Unbalanced threshold 0.65
15000.00
13750.00
12500.00
11250.00
10000.00
Eps
5 31 57 83 109 135 161 187 213 239 266 292 318 344 370 396 422 448 474 500
Min Samples
−0.900 −0.925 −0.950 −0.975 −1.000 −1.025 −1.050 −1.075 −1.100
Silhouette Score

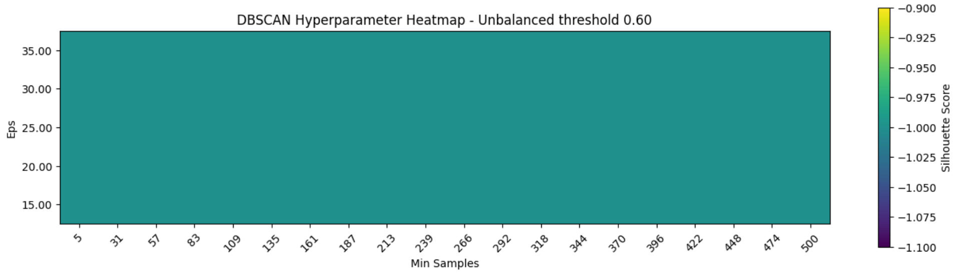

DBSCAN Hyperparameter Heatmap - Unbalanced threshold 0.70
15000.00
13750.00
12500.00
11250.00
10000.00
Eps
5 31 57 83 109 135 161 187 213 239 266 292 318 344 370 396 422 448 474 500
Min Samples
−0.900 −0.925 −0.950 −0.975 −1.000 −1.025 −1.050 −1.075 −1.100
Silhouette Score

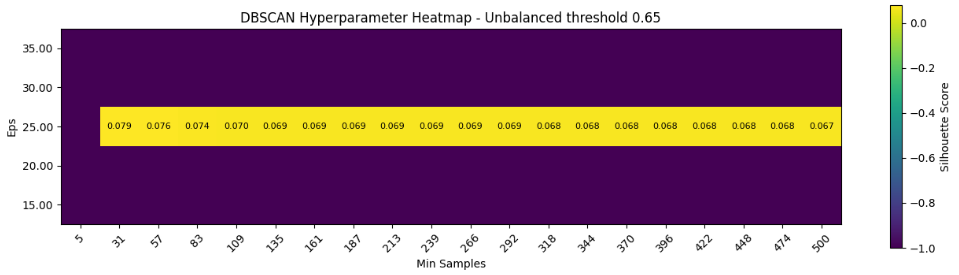

DBSCAN Hyperparameter Heatmap - Unbalanced threshold 0.75
15000.00
13750.00
12500.00
11250.00
10000.00
Eps
5 31 57 83 109 135 161 187 213 239 266 292 318 344 370 396 422 448 474 500
Min Samples
−0.900 −0.925 −0.950 −0.975 −1.000 −1.025 −1.050 −1.075 −1.100
Silhouette Score

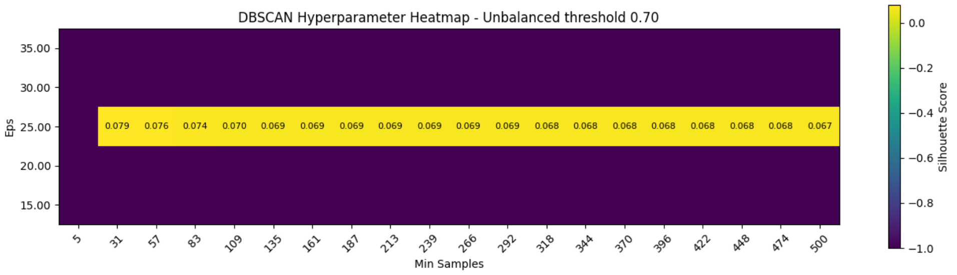
- **VQ-VAE**

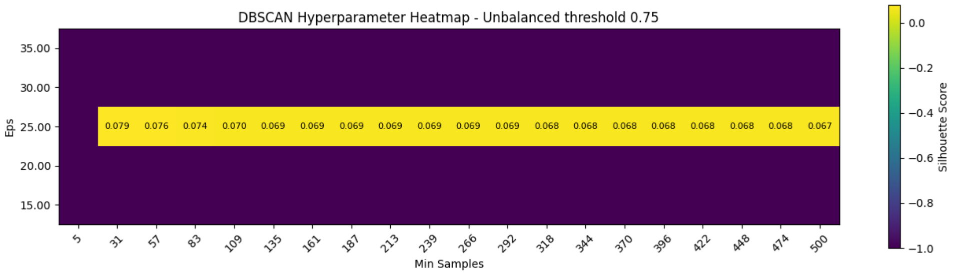

Distances between samples

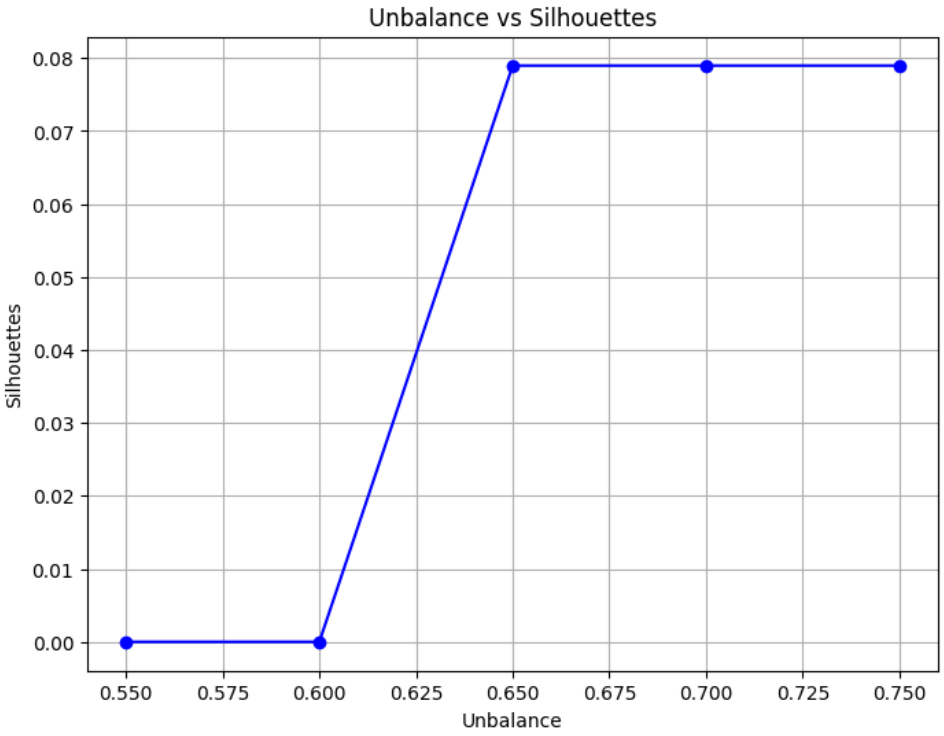

DBSCAN Hyperparameter Heatmap - Unbalanced threshold 0.55

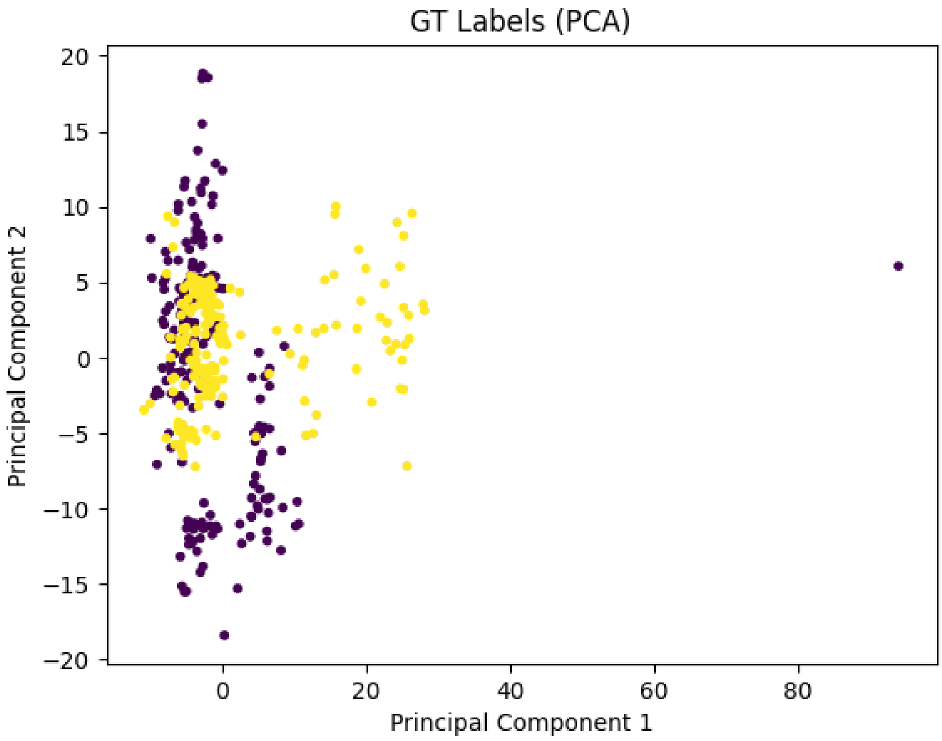

DBSCAN Hyperparameter Heatmap - Unbalanced threshold 0.60


DBSCAN Hyperparameter Heatmap - Unbalanced threshold 0.65

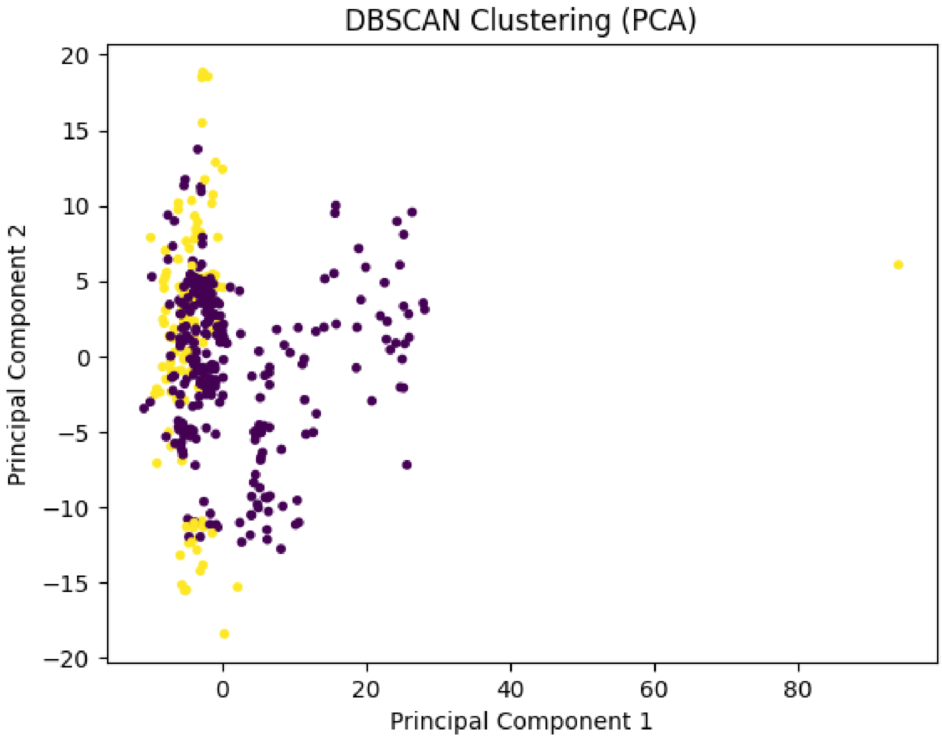

DBSCAN Hyperparameter Heatmap - Unbalanced threshold 0.70
Eps
35.00
30.00
25.00
20.00
15.00
0.079
0.076
0.074
0.070
0.069
0.069
0.069
0.069
0.069
0.069
0.069
0.068
0.068
0.068
0.068
0.068
0.068
0.068
0.067
5
31
57
83
109
135
161
187
213
239
266
292
318
344
370
396
422
448
474
500
Min Samples
Silhouette Score
0.0
−0.2
−0.4
−0.6
−0.8
−1.0

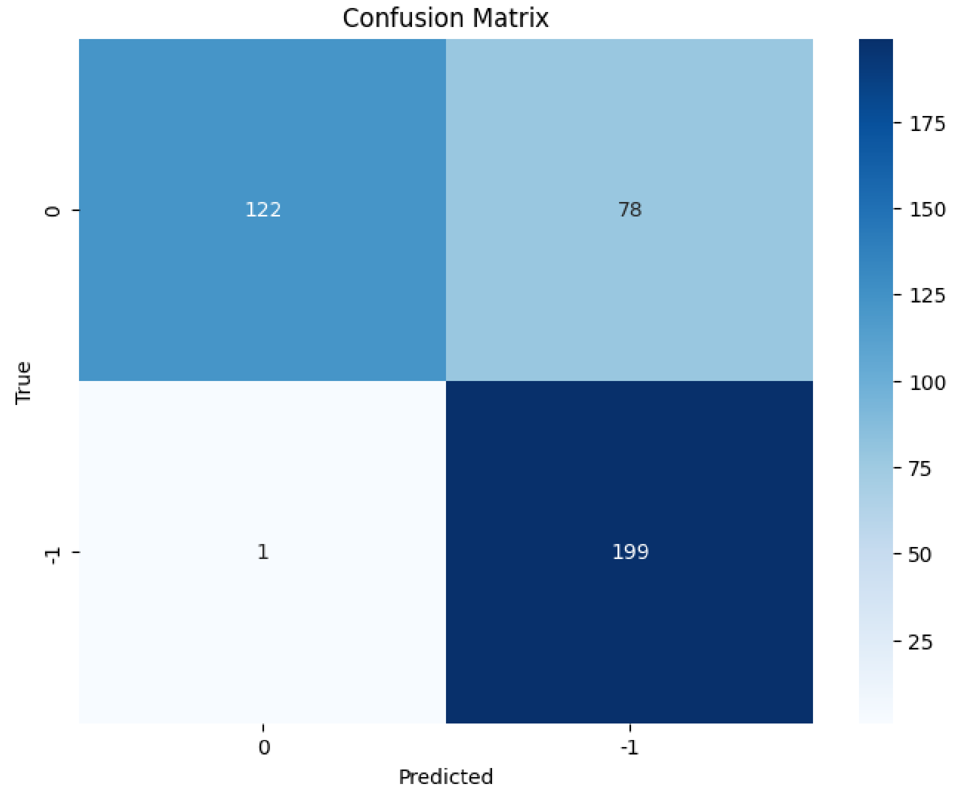

DBSCAN Hyperparameter Heatmap - Unbalanced threshold 0.75
Eps
35.00
30.00
25.00
20.00
15.00
0.079
0.076
0.074
0.070
0.069
0.069
0.069
0.069
0.069
0.069
0.069
0.068
0.068
0.068
0.068
0.068
0.068
0.068
0.067
5
31
57
83
109
135
161
187
213
239
266
292
318
344
370
396
422
448
474
500
Min Samples
Silhouette Score
0.0
−0.2
−0.4
−0.6
−0.8
−1.0


Unbalance vs Silhouettes
Silhouettes
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0.00
0.550
0.575
0.600
0.625
0.650
0.675
0.700
0.725
0.750
Unbalance



- Eps: 25 - Min: 31 - **Threshold 0.65**
  - Silhouette Score: **0.078**
  - Accuracy: **80%**





GMM

In the exploration of Gaussian Mixture Models (GMM), the number of clusters was fixed at 2, aligning with the inherent nature of the dataset that naturally falls into two Gaussian distributions. The choice of covariance type was set to "full," allowing for flexibility in capturing the shape of the clusters, and the initialization parameters were configured to "random_from_data."

- tol (Tolerance): Tolerance represents the convergence threshold for the model. A lower tolerance indicates a more stringent convergence criterion.
- max_iter (Maximum Iterations): Maximum iterations define the cap on the number of iterations the algorithm is allowed to perform before terminating. It helps in preventing infinite loops.

Model Convergence and Initialization Challenges:

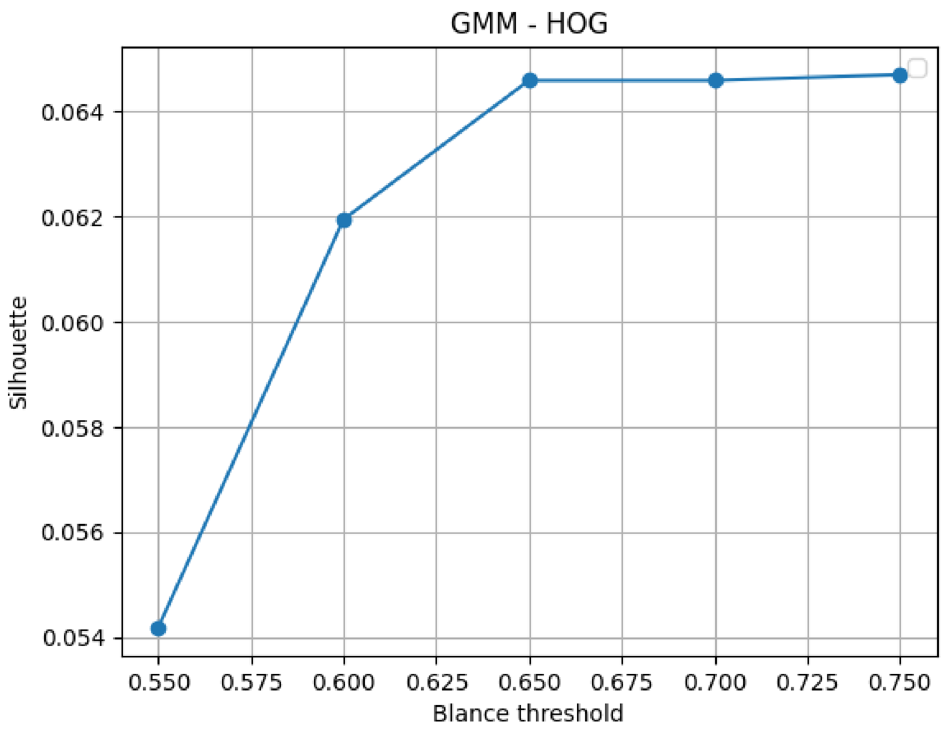
Observations indicated that the GMM model typically converged under a very low tolerance after a minimal number of iterations (1 to 5 iterations). Consequently, changes in tolerance and maximum iterations did not significantly affect the clustering performance, as they were less restrictive than the model's inherent capabilities within the initial iterations.

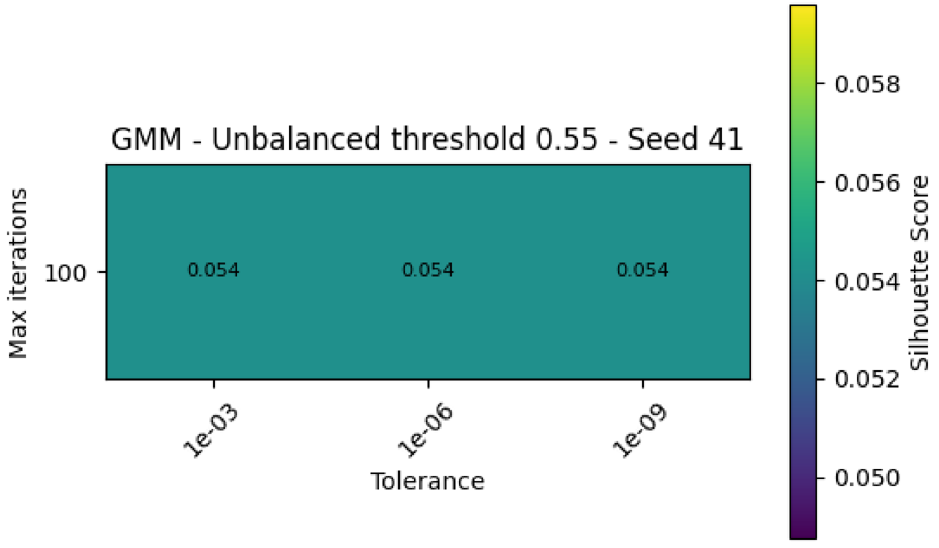
A crucial consideration emerged regarding the structure of the identified Gaussians. While the model successfully captured clusters, the structural integrity of these Gaussians was sensitive to the initializations. This behaviour might be attributed to the ease with which Gaussians can be fitted to the data in just a few steps.

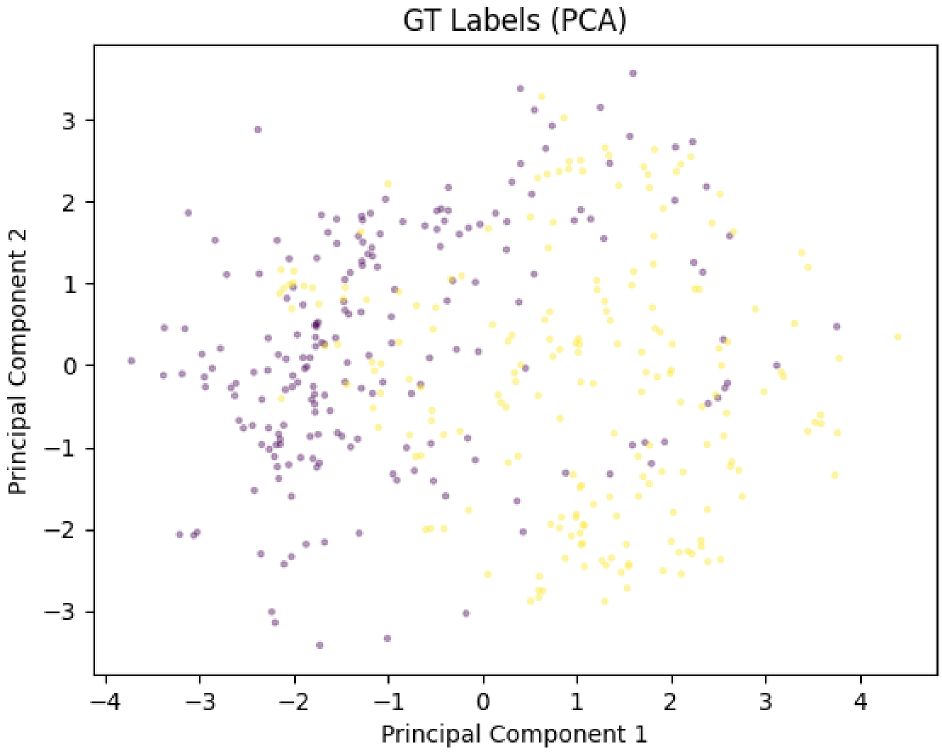
Mitigating Initialization Sensitivity:

To address the initialization sensitivity, a grid search was conducted for 100 seed values, ensuring a thorough exploration of potential initializations. The models with the highest Silhouette Scores under the defined imbalance constraints were retained for further analysis, ensuring robustness in capturing meaningful cluster structures within the vegetable dataset.

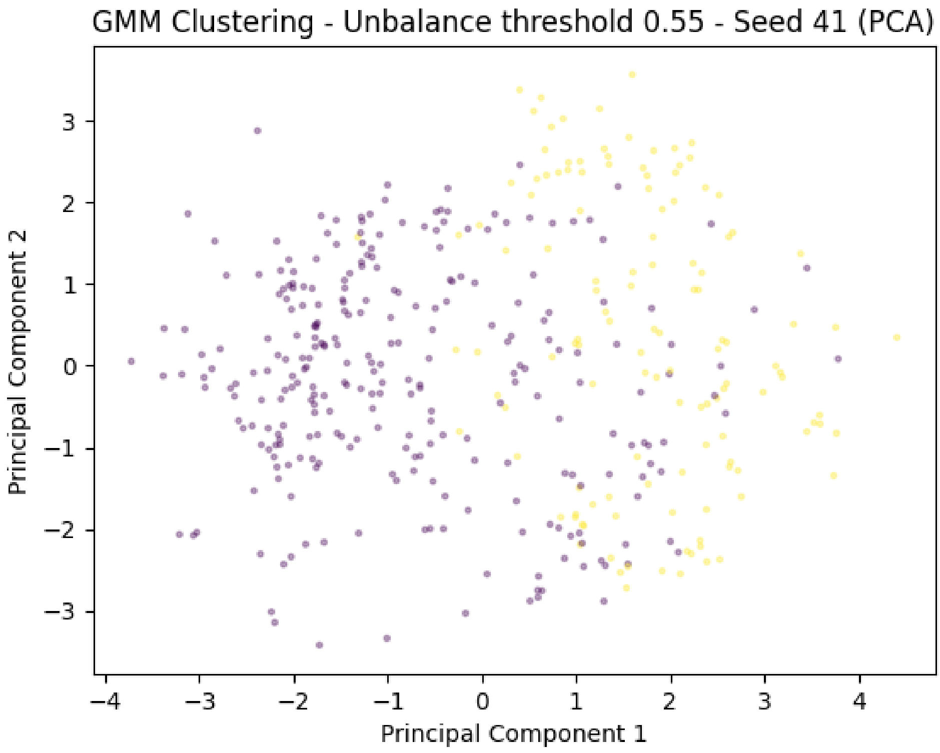
- HOG

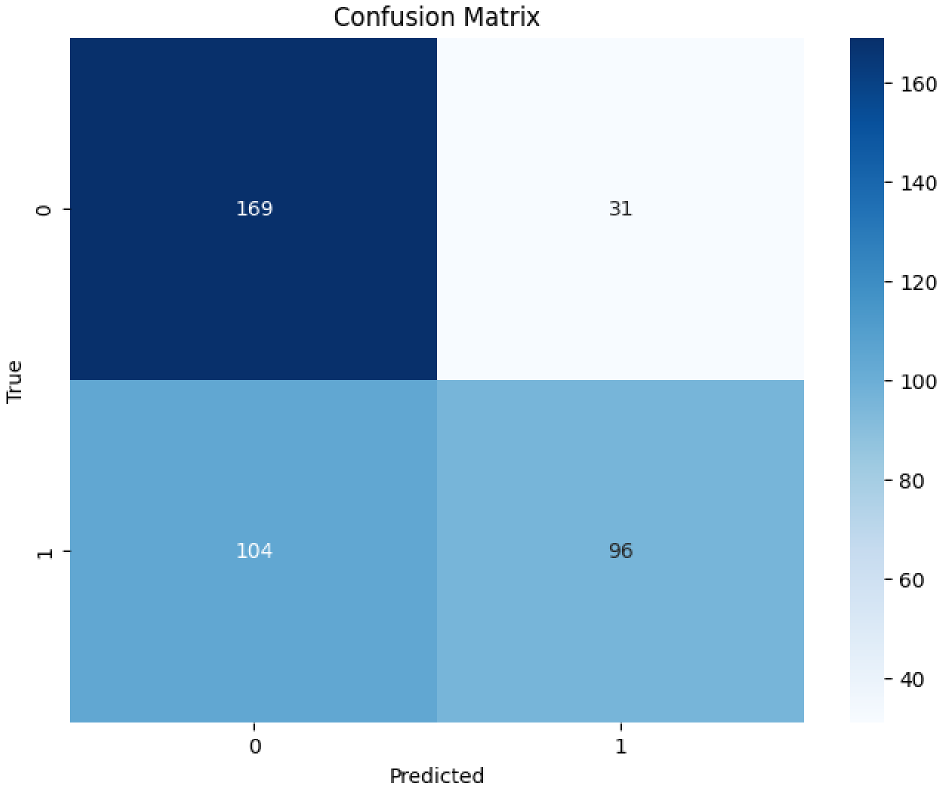




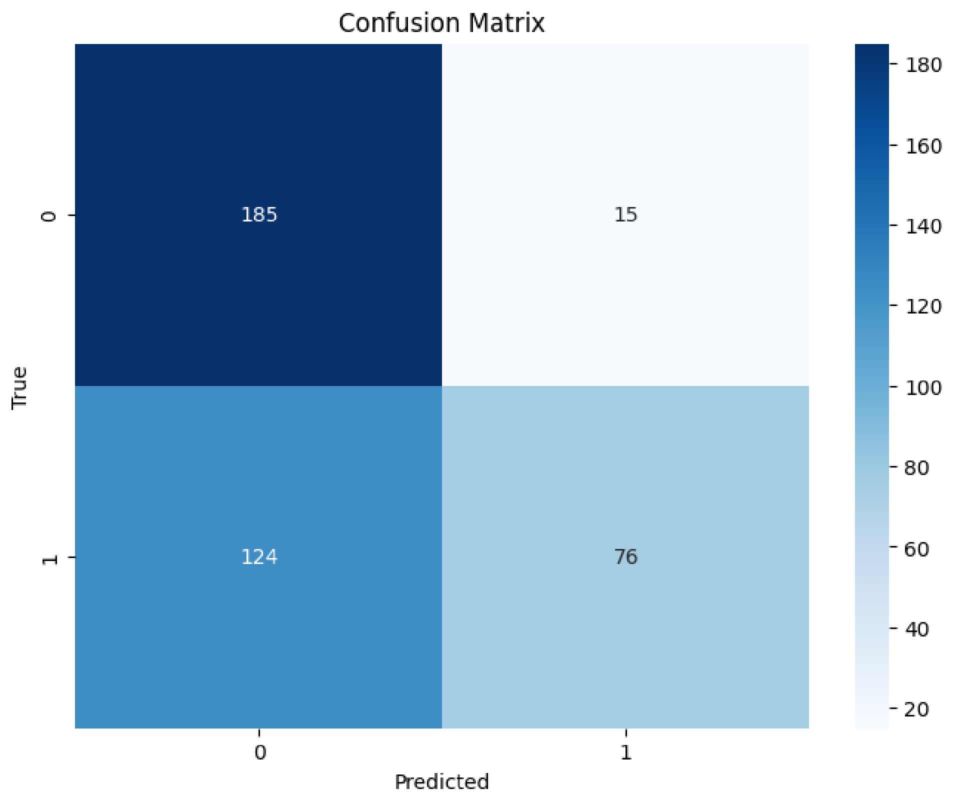


- Seed: 41 - **Threshold 0.55**
  - Silhouette Score: **0.054**
  - Accuracy: **66%**

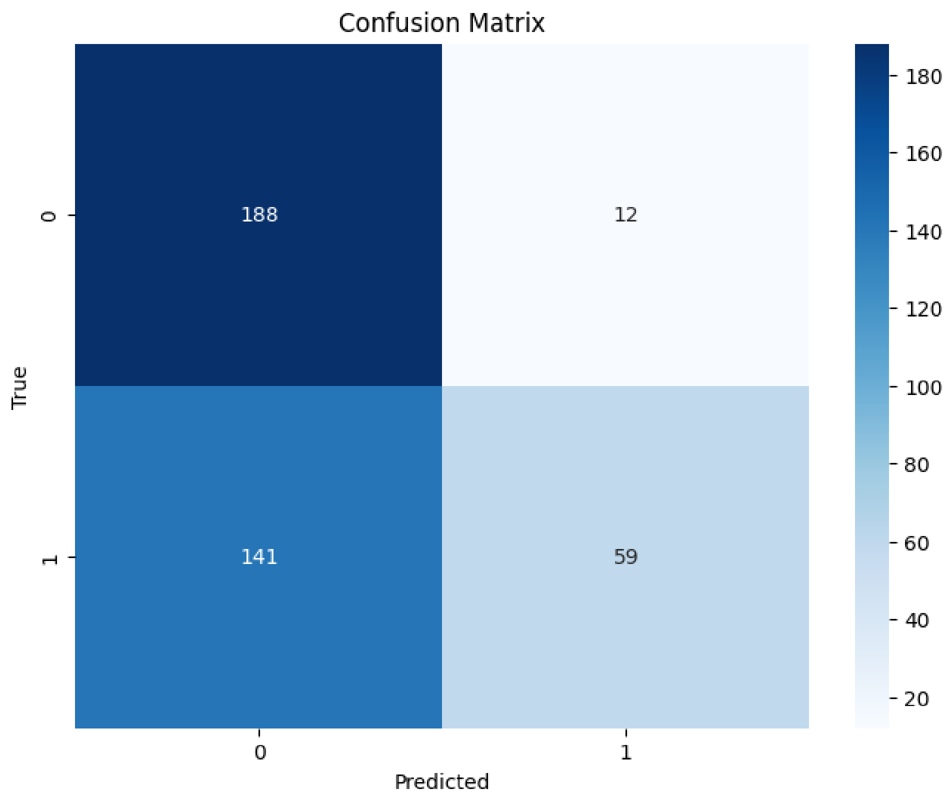




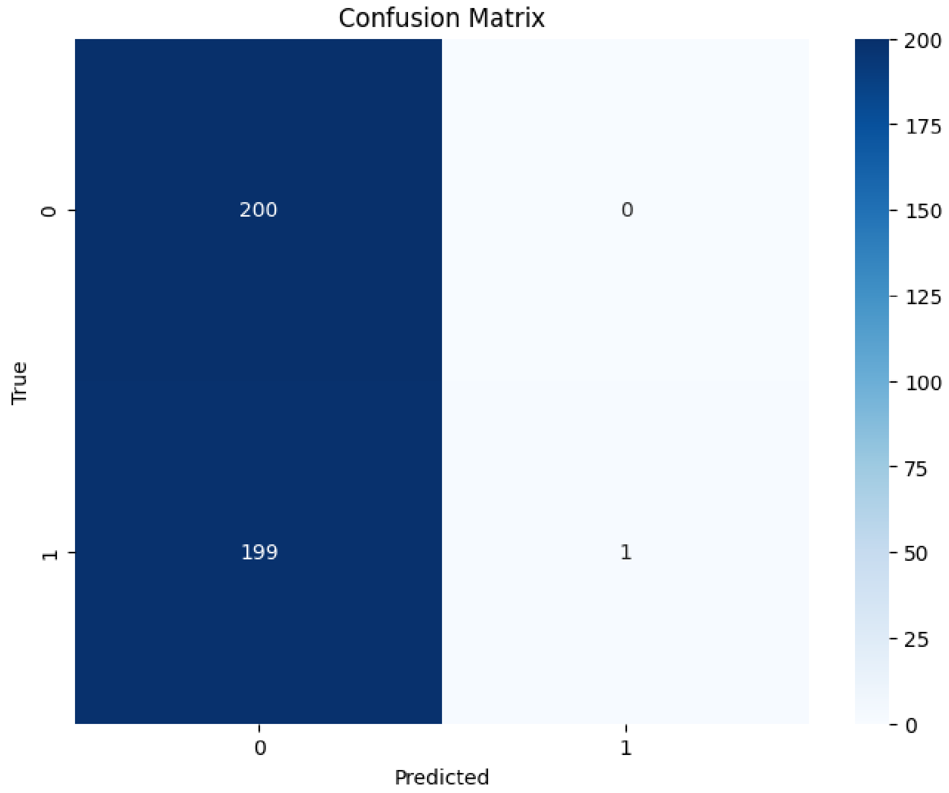
- Seed: 21 - **Threshold 0.6**
  - Silhouette Score: **0.062**
  - Accuracy: **65%**



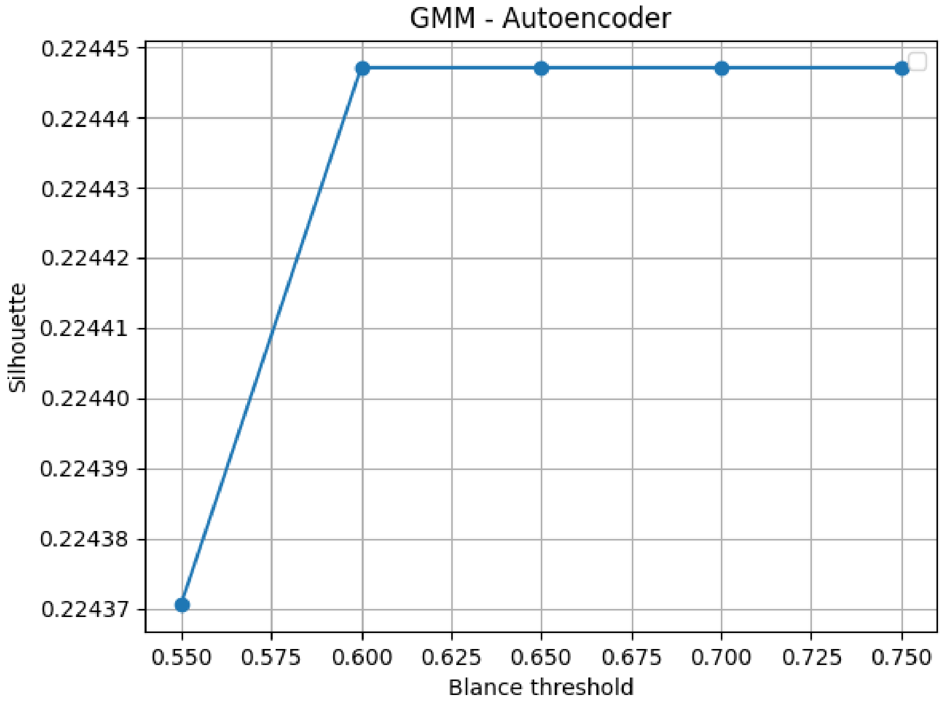
- Seed: 73 - **Threshold 0.65**
    - Silhouette Score: **0.064**
    - Accuracy: **62%**
- Seed: 73 - **Threshold 0.70**
    - Silhouette Score: **0.064**
    - Accuracy: **62%**

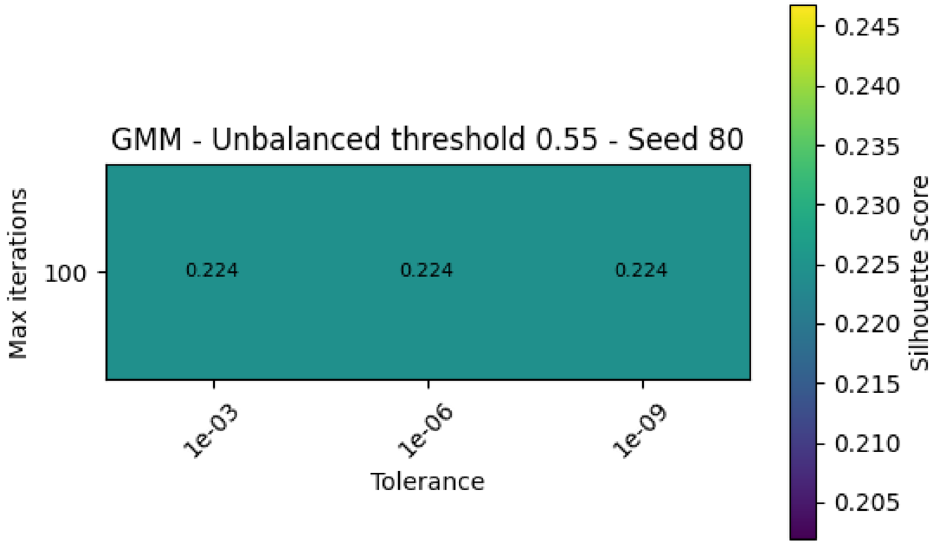


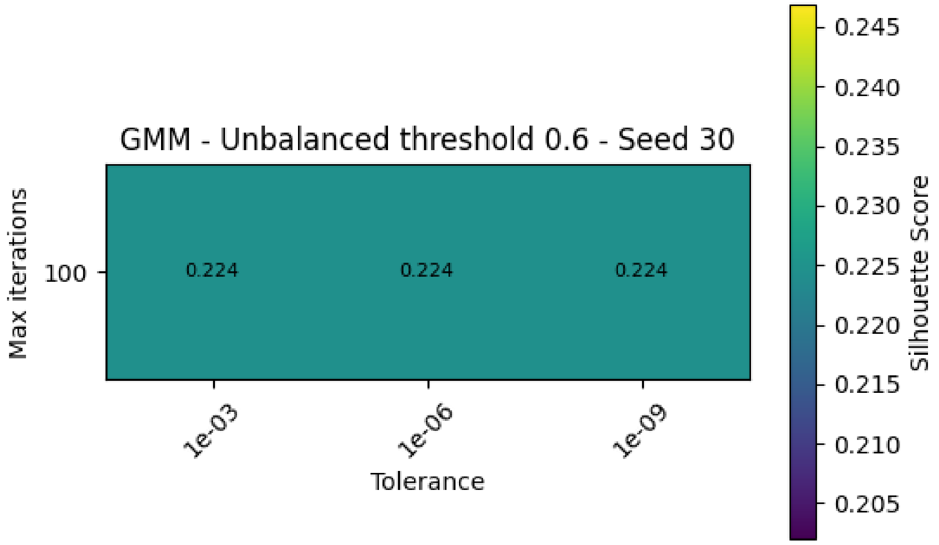
- Seed: 55 - **Threshold 0.75**
  - Silhouette Score: **0.064**
  - Accuracy: **50%**

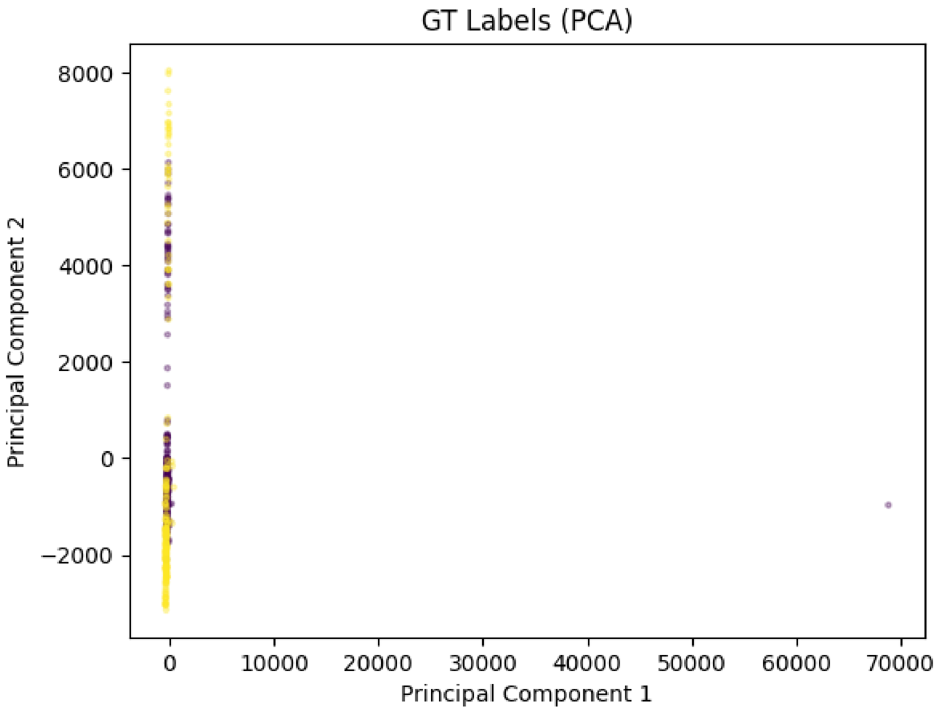


- RGB Histogram

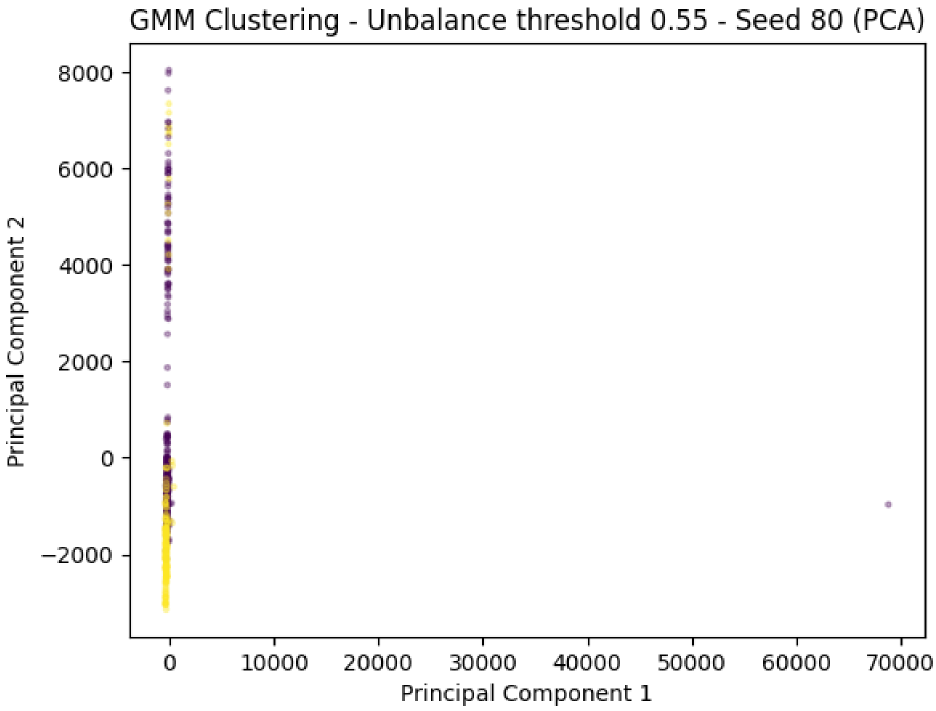


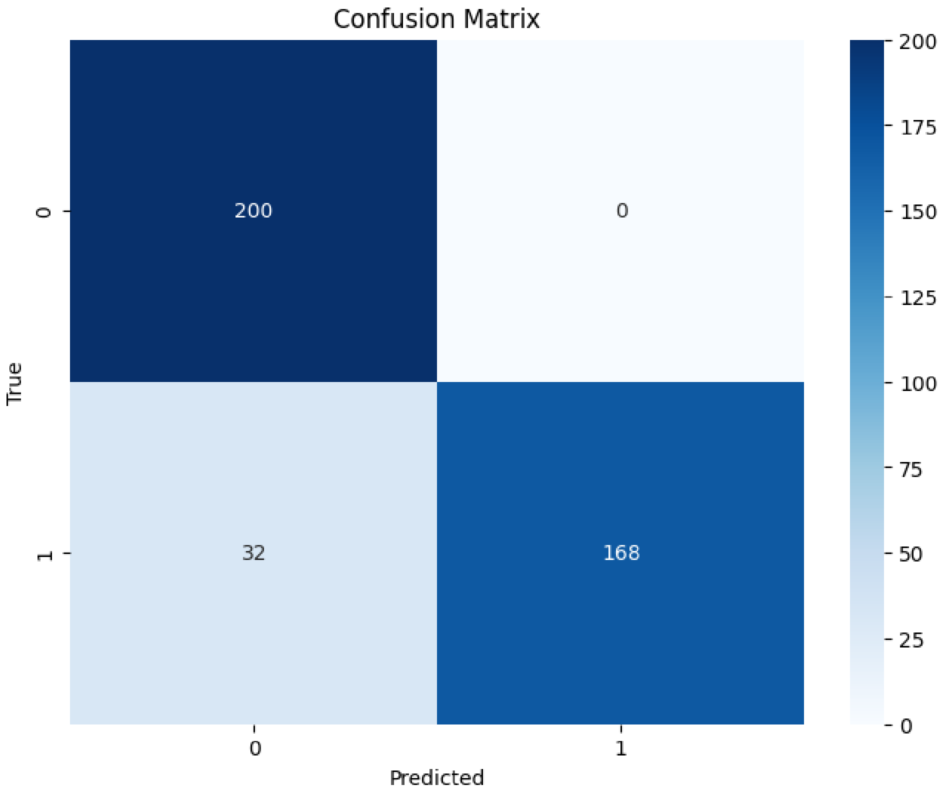




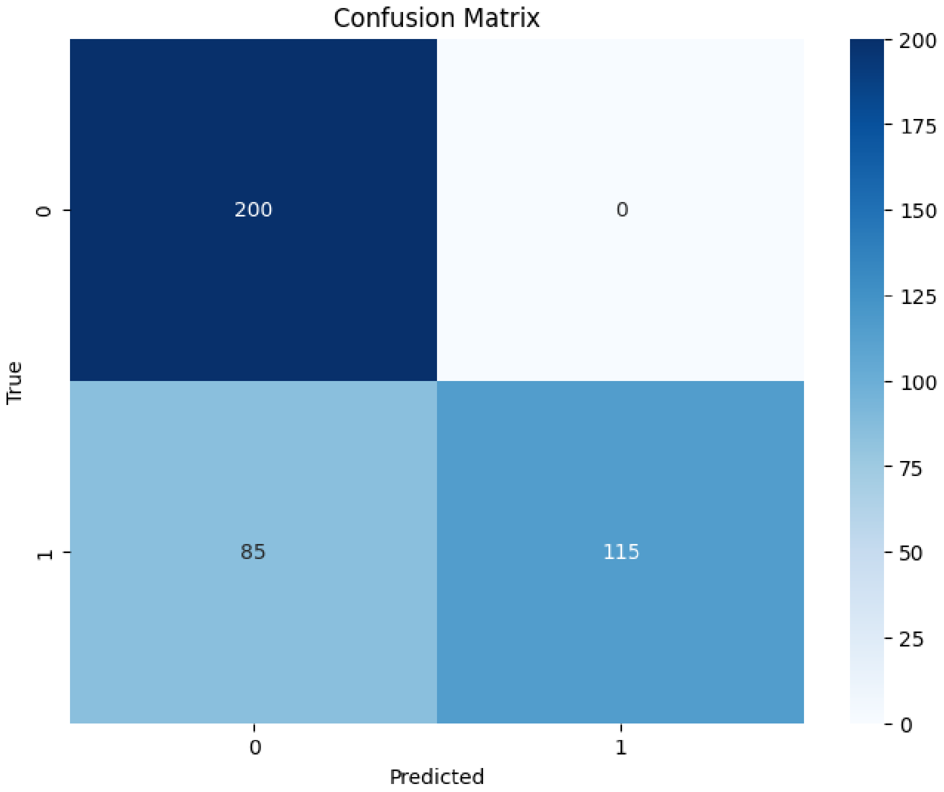


- Seed: 80 - **Threshold 0.55**
  - Silhouette Score: **0.2243**
  - Accuracy: **92%**

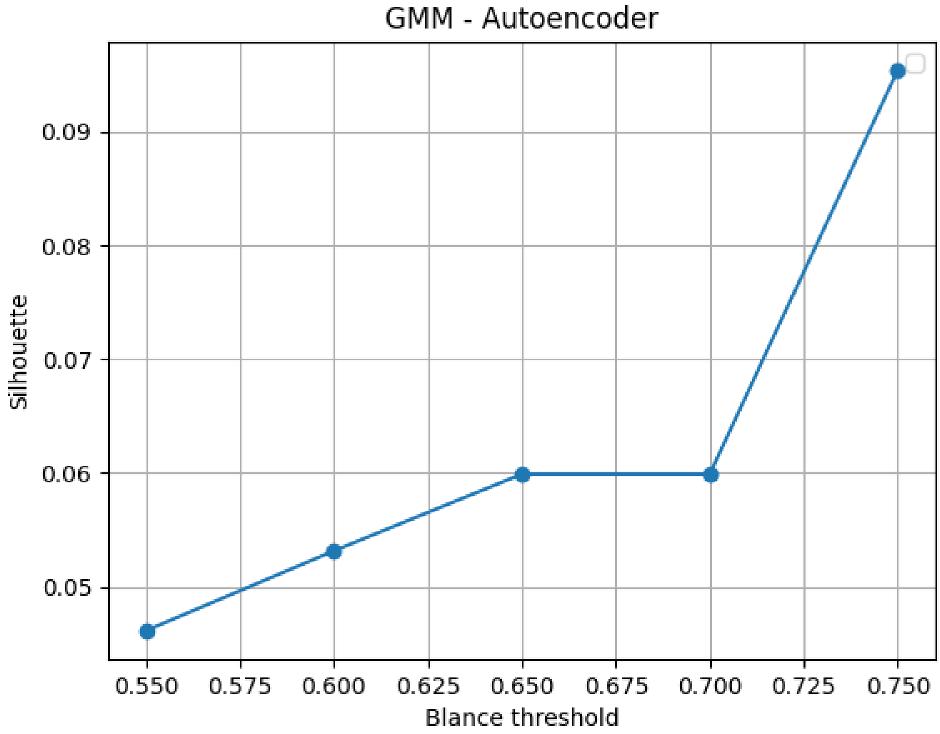


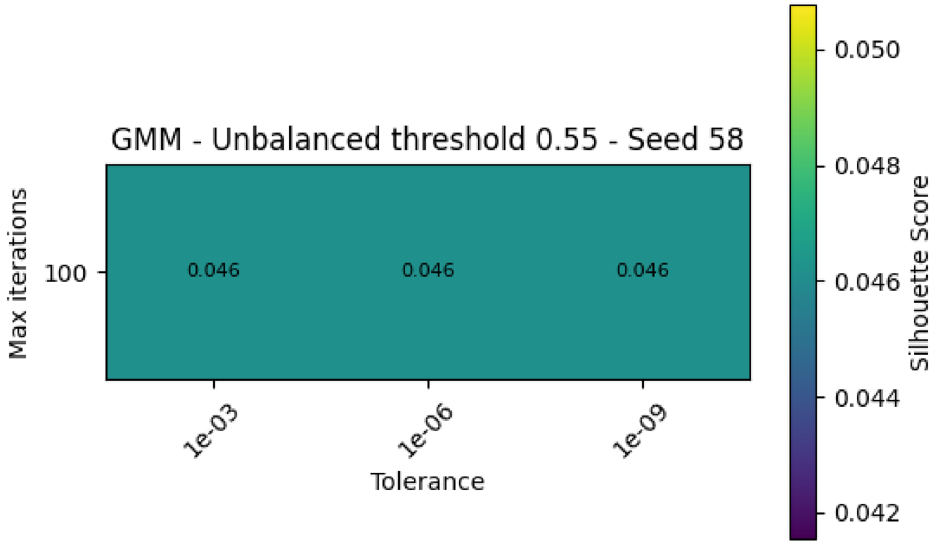


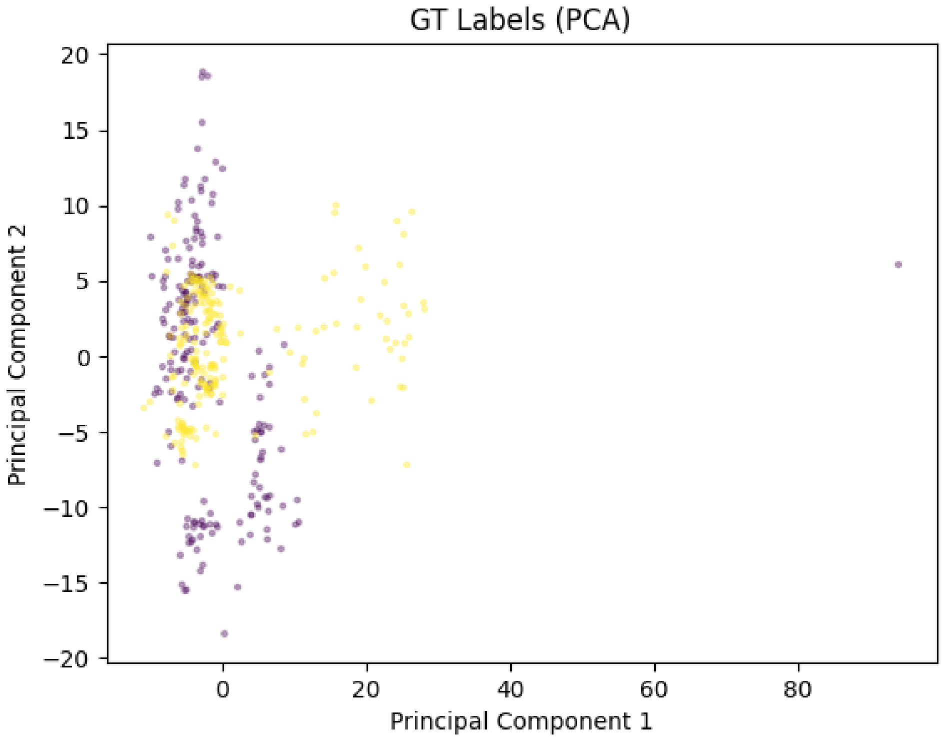
- Seed: 30 - **Threshold 0.6**
  - Silhouette Score: **0.2244**
  - Accuracy: **79%**


Confusion Matrix
True
0
1
200
0
85
115
0
1
Predicted
200
175
150
125
100
75
50
25
0

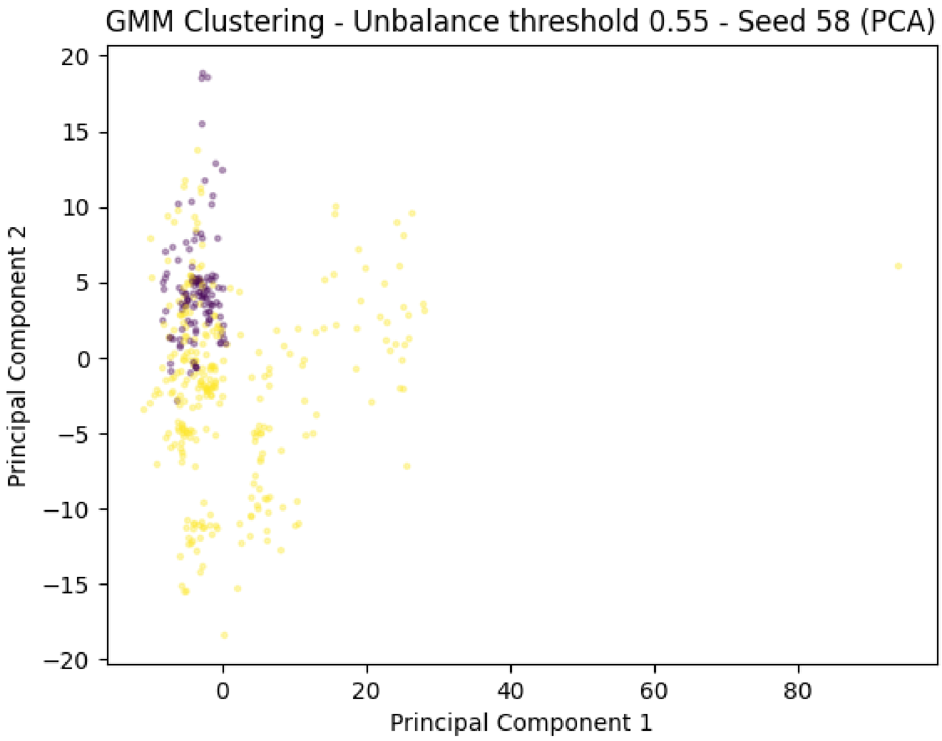
- VQ-VAE

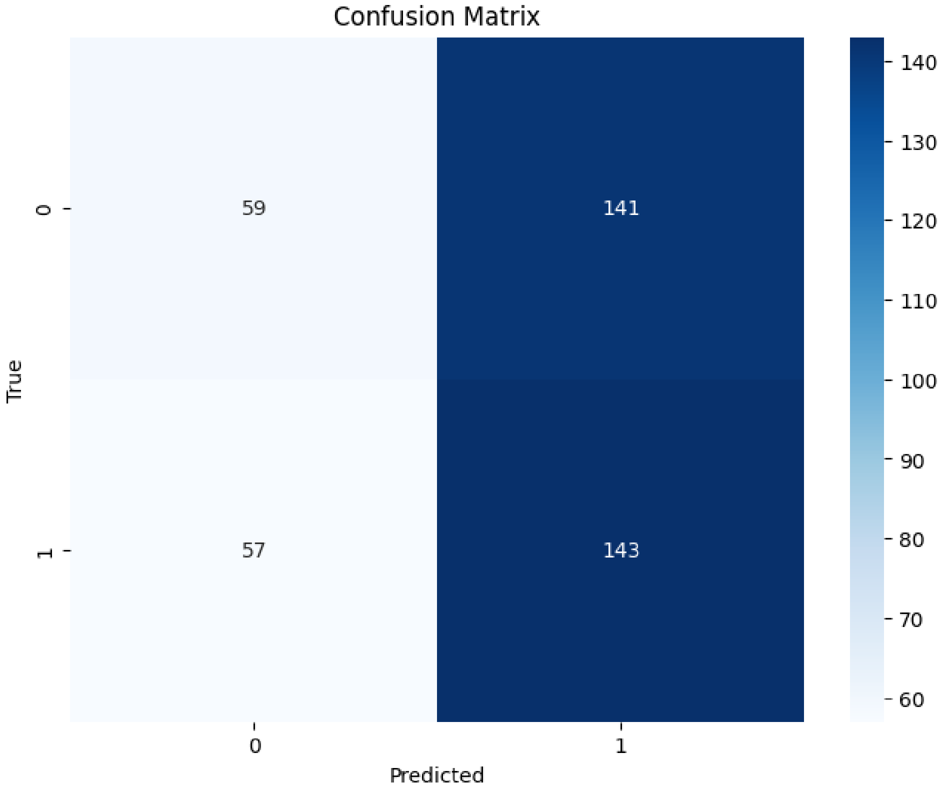

GMM - Autoencoder
Silhouette
0.09
0.08
0.07
0.06
0.05
0.550
0.575
0.600
0.625
0.650
0.675
0.700
0.725
0.750
Blance threshold



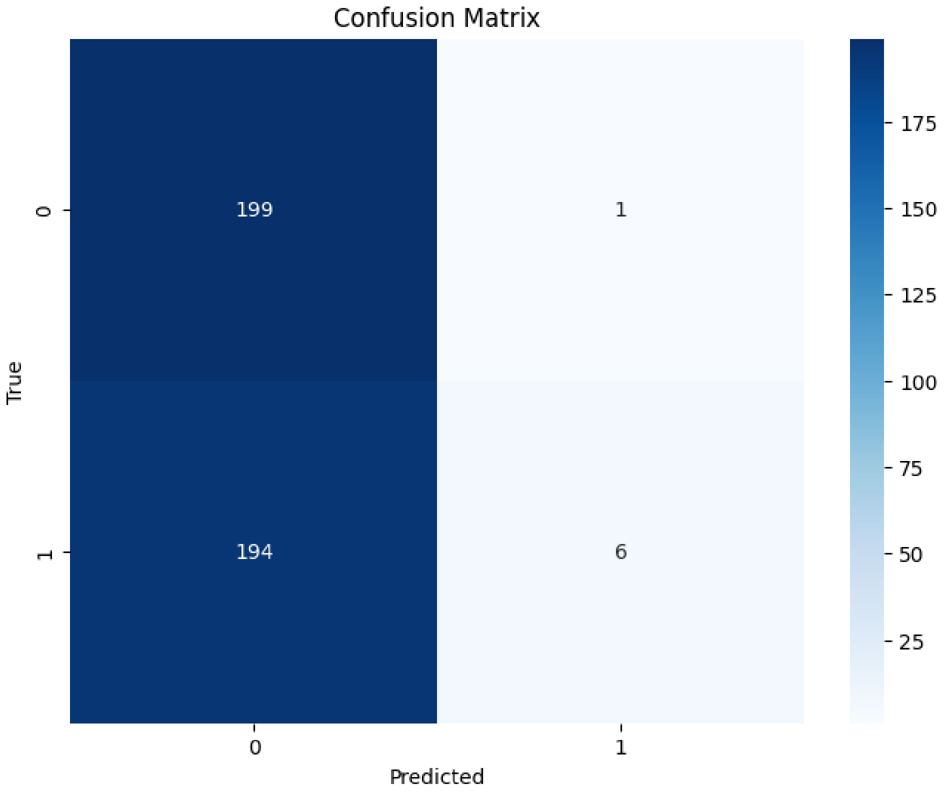


- Seed: 58 - **Threshold 0.55**
  - Silhouette Score: **0.046**
  - Accuracy: **51%**

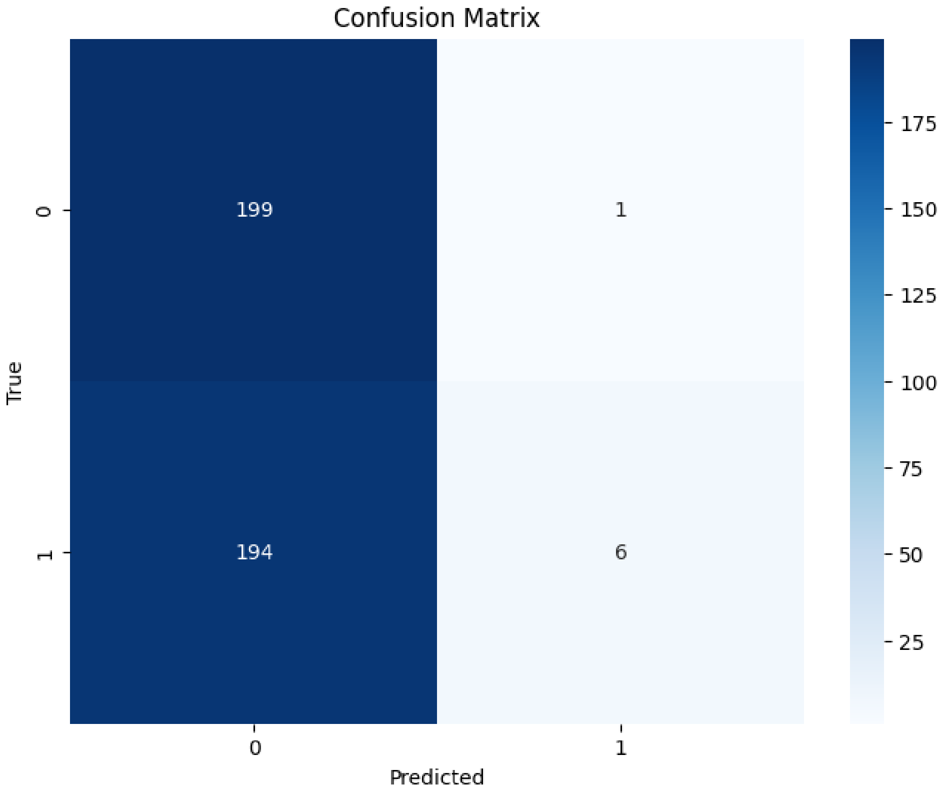




- Seed: 89 - **Threshold 0.6**
  - Silhouette Score: **0.06**
  - Accuracy: **51%**



- Seed: 89 - **Threshold 0.65**
  - Silhouette Score: **0.06**
  - Accuracy: **50%**



Results and Analysis - Performance Metrics

	SVM	RF	DBSCAN	GMM	*Random*
HOG	98%	92%	68% [0.046]	66% [0.054]	*50%*
RGB Histogram	100%	100%	-	**92% [0.224]**	*50%*
VQ-VAE	99%	99%	**80% [0.078]**	51% [0.046]	*50%*