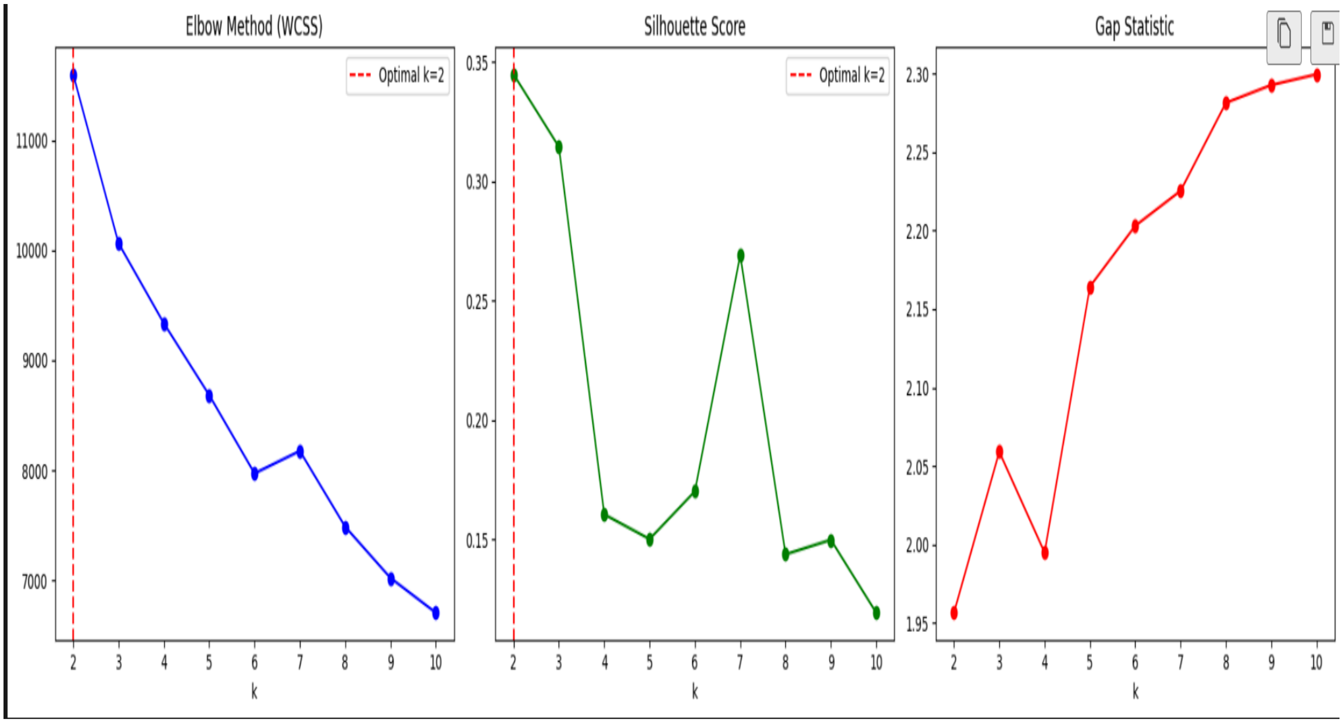
LAB 4: Clustering and Dimensionality Reduction Analysis

Part 1: Baseline Clustering (Original Data)

Experiment 1: K-Means Clustering

We first determined the optimal number of clusters (k) using the Elbow Method, Silhouette Score, and Gap Statistic.

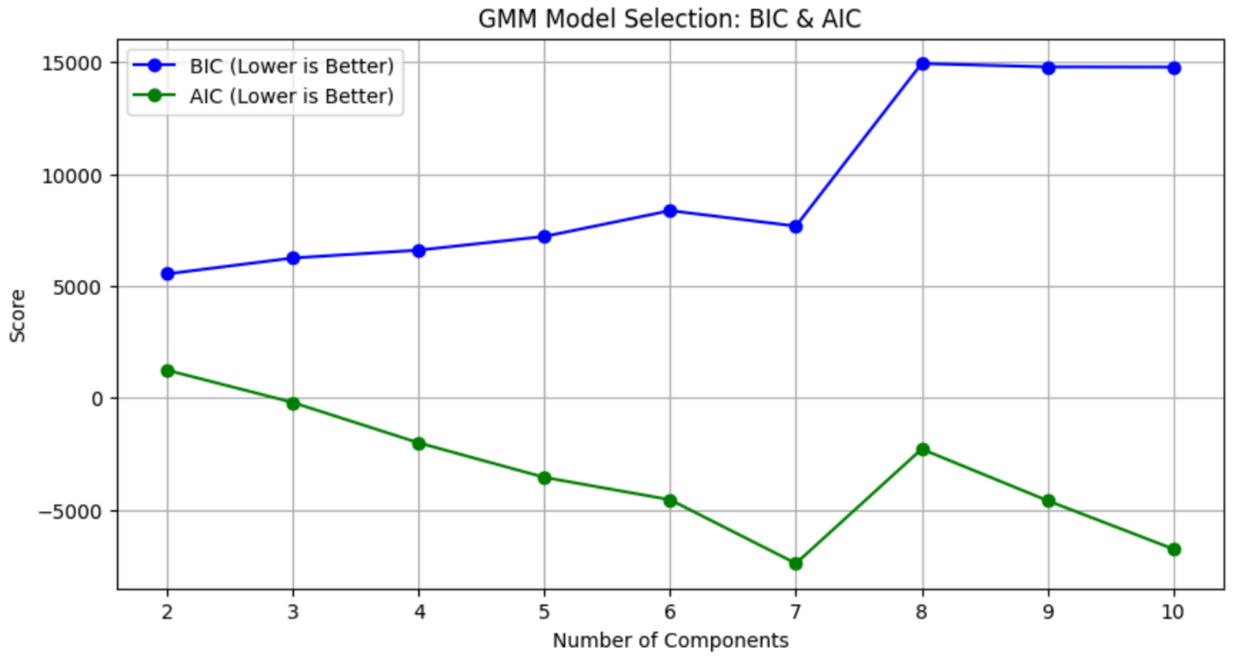
- **Optimal k:** All three metrics agreed on **k=2**, which aligns with the ground truth (Malignant vs. Benign).
- **Initialization:** K-Means++ converged faster (12 iterations) compared to random initialization, showing more stable inertia reduction.
- **Performance:**
  - **Purity:** 0.9051
  - **ARI:** 0.6536
  - **Silhouette Score:** 0.3434



Experiment 2: Gaussian Mixture Models (GMM)

We tested GMM with cluster counts from 2 to 10 and four covariance types.

- **Model Selection:** The BIC score was lowest at **k=2**, confirming the optimal cluster count.
- **Covariance Analysis:** The **'full'** covariance type yielded the highest Log-Likelihood (372) compared to 'tied', 'diagonal', and 'spherical', indicating that the features have complex correlations that simpler models fail to capture.
- **Performance (k=2, full):**
  - **Purity:**0.9332 (Comparable to K-Means)
  - **ARI:** 0.7495

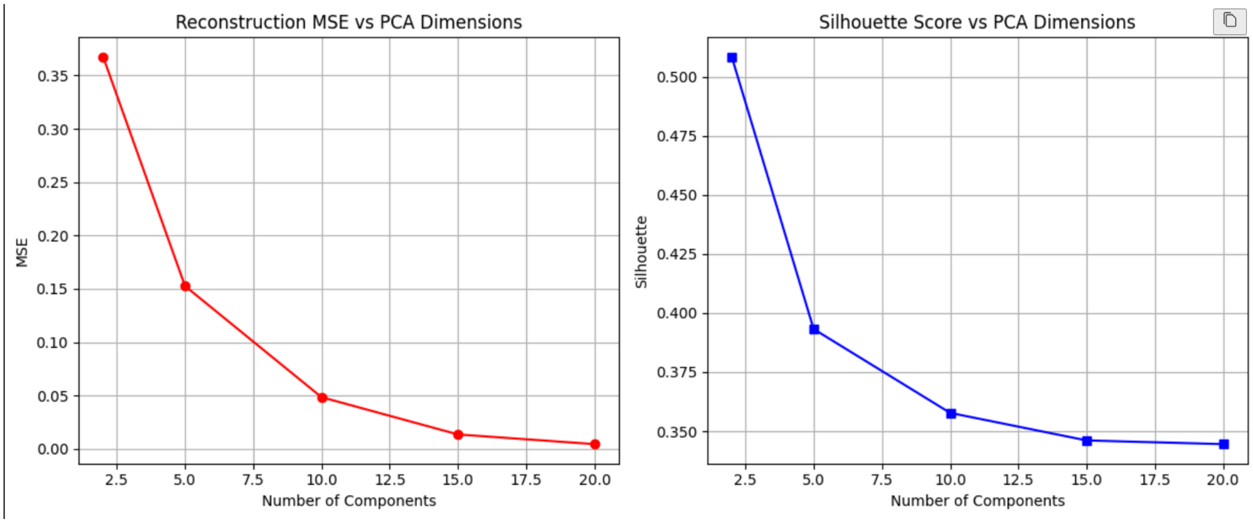


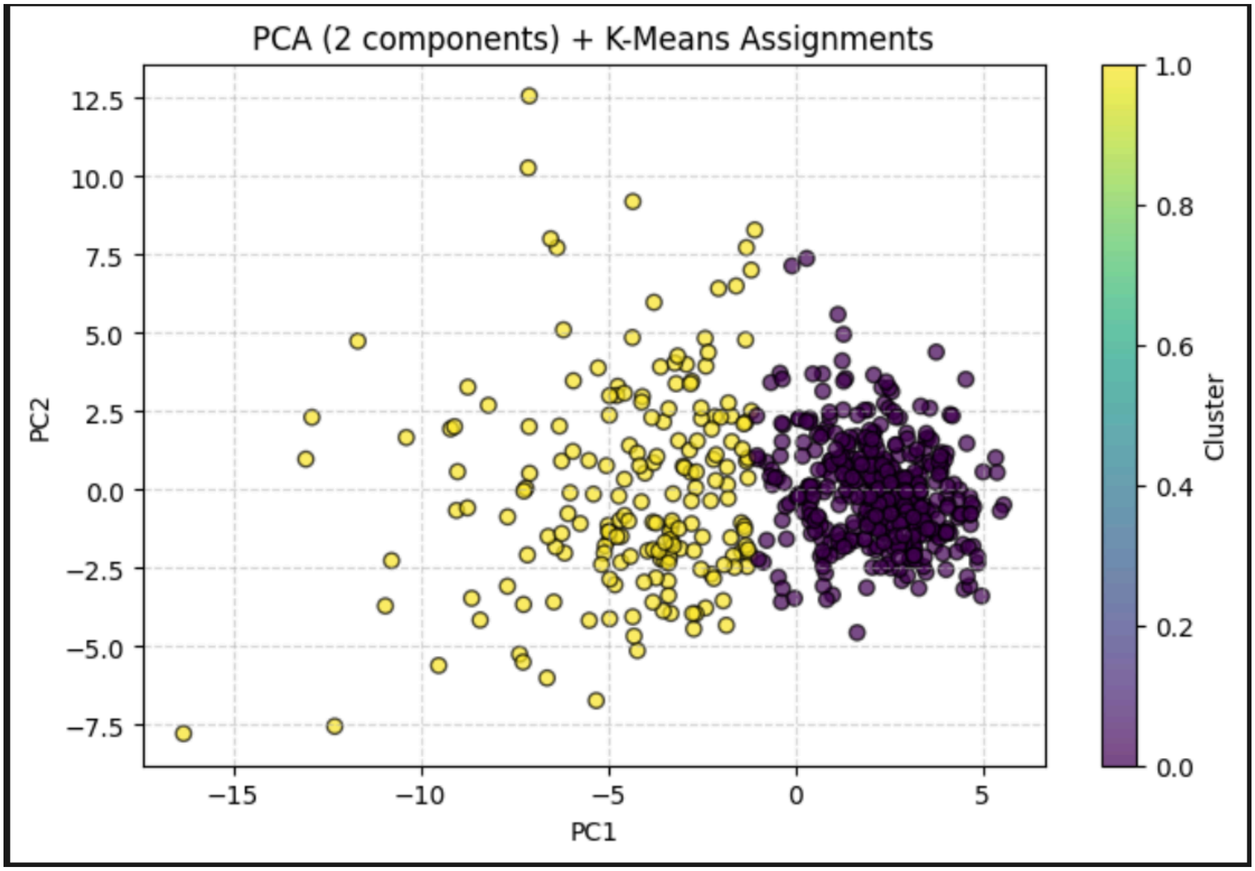
3. Part 2: Dimensionality Reduction Analysis

Experiment 3: K-Means after PCA

We evaluated K-Means performance across PCA dimensions [2, 5, 10, 15, 20].

- **Reconstruction vs. Clustering Trade-off:**
  - As expected, **Reconstruction MSE** decreases as dimensions increase (from 0.36 at 2D to 0.004 at 20D).
  - However, the **Silhouette Score** is highest at **2 Dimensions** (0.508) and drops significantly as dimensions increase. This suggests that while higher dimensions capture more variance, they introduce sparsity that blurs cluster boundaries.
- **Best Performance:**
  - The highest **Purity (0.9086)** was achieved at **Dimensions 2** , outperforming the baseline K-Means (0.905).
  - This indicates PCA effectively denoises the data, improving cluster purity.

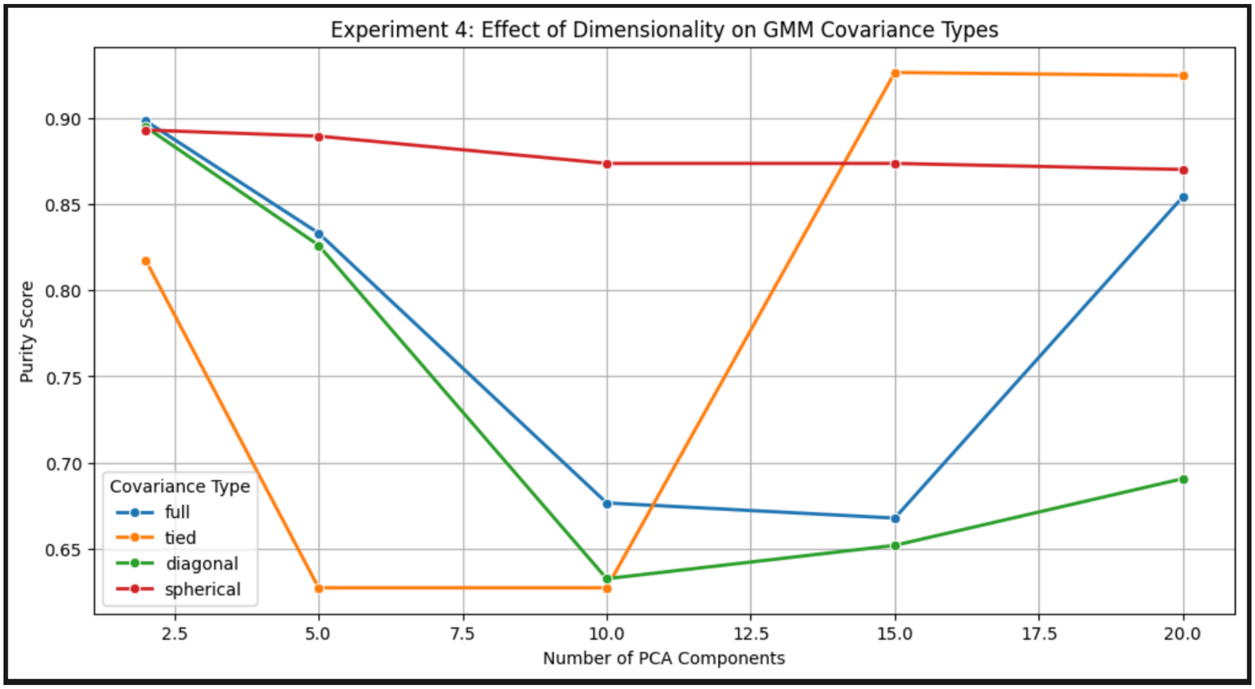




Experiment 4: GMM after PCA

We analyzed how dimensionality impacts the optimal covariance type.

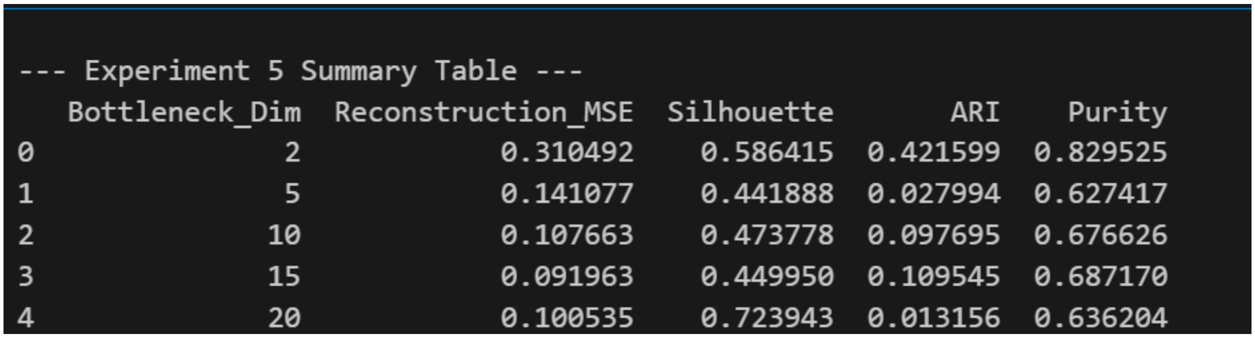
- **Impact of Dimensionality:**
    - **Full Covariance:** Performed well at lower dimensions (d=2, 5) but performance degraded or became unstable at higher dimensions (d=20) due to the large number of parameters to estimate.
    - **Tied Covariance:** Achieved the overall **Best Configuration** at **PCA=15** with a Purity of **0.926**. Tied covariance works well in higher dimensions as it shares parameters across clusters, reducing overfitting.
    - **Diagonal/Spherical:** Generally underperformed, proving that feature correlations are significant in this dataset.



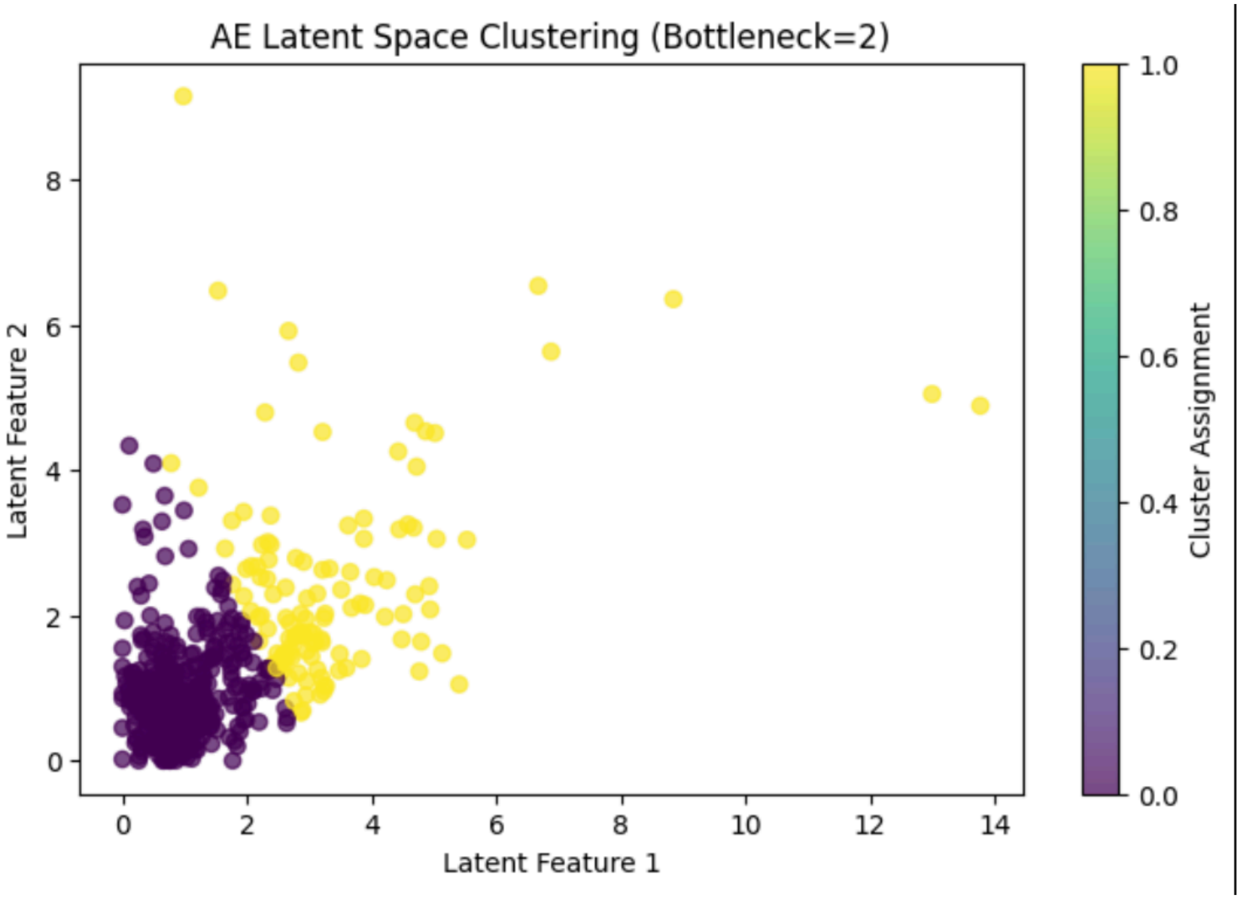
Experiment 5: K-Means after Autoencoder

We trained an Autoencoder with bottleneck dimensions [2, 5, 10, 15, 20] and applied K-Means to the latent space.

- **Latent Space Analysis:**
  - **Reconstruction MSE** dropped from 0.31 (Dim 2) to 0.10 (Dim 20), confirming the Autoencoder successfully learned to compress and reconstruct the data.
  - **Clustering Performance:** The best **Purity (0.829)** was observed at **Bottleneck=2**. However, performance was notably unstable at higher dimensions (e.g., Purity dropped to ~0.44 at Dim 5), suggesting the Autoencoder's latent space might be encoding features useful for reconstruction but not necessarily for separation (malignant vs benign) without supervised guidance.
- **Comparison:** Autoencoder-based K-Means generally underperformed compared to PCA-based K-Means (Max Purity 0.829 vs 0.912).



```
--- Experiment 5 Summary Table ---
   Bottleneck_Dim  Reconstruction_MSE  Silhouette       ARI    Purity
0               2            0.310492    0.586415  0.421599  0.829525
1               5            0.141077    0.441888  0.027994  0.627417
2              10            0.107663    0.473778  0.097695  0.676626
3              15            0.091963    0.449950  0.109545  0.687170
4              20            0.100535    0.723943  0.013156  0.636204
```

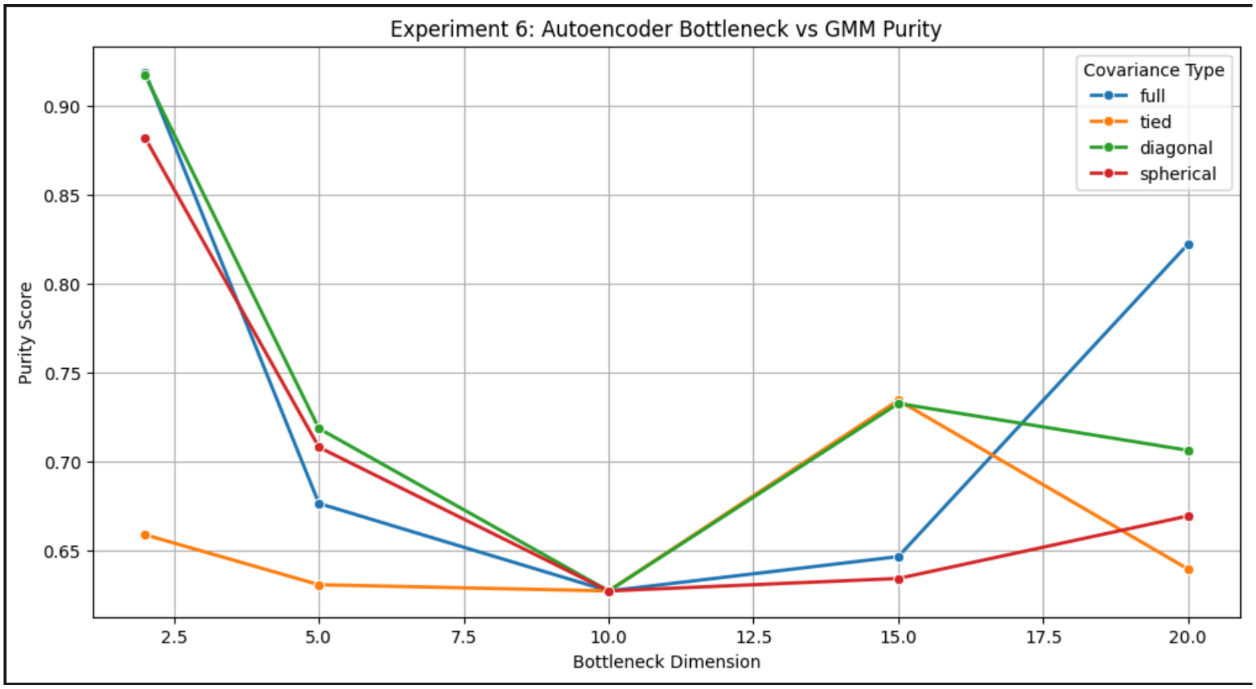


Experiment 6: GMM after Autoencoder

We applied GMM to the Autoencoder's latent features.

- **Results:**
  - Unlike K-Means, GMM was able to exploit the non-linear structure of the AE latent space effectively.
  - **Best Configuration: Latent Dim=2** with **Full Covariance** achieved a Purity of **0.919**, which is comparable to the best PCA results.

- ○ This highlights that while the AE latent space might not form spherical clusters (bad for K-Means), it forms complex densities that GMM can model well.



4. Comprehensive Comparison & Conclusion

The following table summarizes the best results from each method:

Method	**Best Config**	**Purity**	**ARI**	**Silhouette**
K-Means (Original)	k=2	0.9051	0.6536	0.3434
GMM (Original)	k=2, Full	0.9033	0.6495	0.2932
PCA + K-Means	**Dim=2**	**0.9121**	**0.6765**	**0.5080**
PCA + GMM	**Dim=15, Tied**	**0.9262**	**0.7232**	0.3352
AE + K-Means	Dim=2	0.8295	0.3105	0.5864
AE + GMM	Dim=2, Full	0.9192	0.7008	0.5025

Our Insights:

1. **Dimensionality Reduction Helps:** Both PCA and Autoencoders (with GMM) achieved higher purity scores than clustering on the raw data. PCA was particularly robust, providing the most consistent improvements.
2. **PCA vs. Autoencoder:** PCA outperformed the Autoencoder for K-Means clustering. The linear separation maximized by PCA aligns well with the diagnosis labels. The Autoencoder achieved high Silhouette scores (tight clusters) but lower Purity in some configurations, indicating it formed clusters based on features not relevant to the diagnosis.
3. **GMM vs. K-Means:** GMM combined with PCA (Tied Covariance) achieved the **highest overall accuracy (92.6%)**. The ability of GMM to model elliptical clusters and tied covariance to handle higher dimensions without overfitting proved to be the winning combination.