UNSUPERVISED CLUSTERING FOR BREAST CANCER TUMOR CLASSIFICATION: A REPRODUCTION STUDY

1. Introduction

Breast cancer is the most common cancer among women worldwide, accounting for about 25% of all female cancer cases, it's the major health problem in the united states . Early detection and accurate diagnosis are paramount in improving patient prognosis and survival rates. Machine learning (ML) methodologies have demonstrated considerable promise in automating and enhancing diagnostic precision by analyzing tumor characteristics obtained from medical imaging and biopsies. This study employs the Breast Cancer Wisconsin (Diagnostic) dataset, comprising 569 patient samples characterized by 30 morphological and textural features derived from digitized images of fine-needle aspirate biopsies. Each sample is classified as benign (n=357) or malignant (n=212). This project aims to evaluate the efficacity of unsupervised clustering algorithms in delineating intrinsic data structures reflective of malignancy status, contrasted against supervised classification using Support Vector Machines (SVM).

2. Literature Review

Recent research has shown the growing importance of machine learning techniques in cancer diagnosis. Zhang and Wang (2023) demonstrated that different clustering algorithms reveal varying biological structures in cancer datasets, emphasizing the need for careful algorithm selection. Similarly, Zhang, Yoon, and Lam (2013) combined K-means clustering with Support Vector Machines (SVM) for breast cancer diagnosis, showing that unsupervised clustering can enhance classification performance.

Gal et al. (2020) highlighted how unsupervised methods can uncover meaningful patterns in breast cancer metabolomics data, supporting the value of clustering in clinical research. Alirezazadeh et al. (2018) further showed that representation learning improves clustering outcomes on complex medical images. In parallel, La Moglia and Almustafa (2024) confirmed that supervised models like SVM consistently achieve high diagnostic accuracy in breast cancer prediction.

Inspired by these studies, our project aims to evaluate different clustering methods and benchmark them against a supervised SVM classifier on the Breast Cancer Wisconsin dataset.

3. Methodology

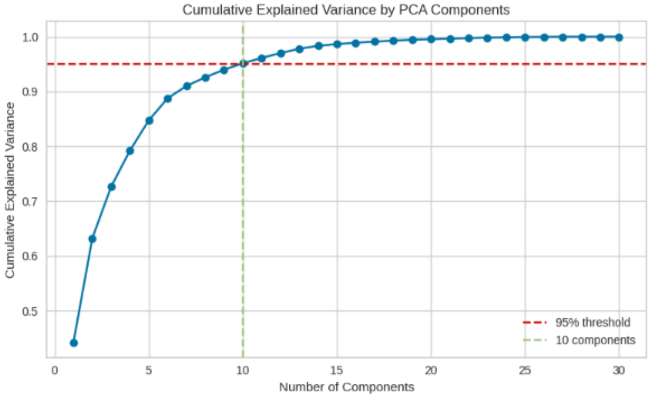
The proposed methodology for this project follows a structured pipeline starting from data acquisition to classification. We utilize the Wisconsin Diagnostic Breast Cancer (WDBC) dataset, which comprises 569 instances with 30 numerical features derived from cell nucleus characteristics such as radius, texture, perimeter, area, smoothness, concavity, and symmetry. Each of these features includes measurements of the mean, standard error, and worst (maximum) value. Since the features vary in scale, we applied standard normalization to ensure consistent distance calculations during clustering. After preprocessing, we perform Principal Component Analysis (PCA) for dimensionality reduction, retaining 95% of the variance, which simplifies the dataset while preserving its essential structure. This reduced feature set is then passed into various unsupervised clustering algorithms, including K-Means, Gaussian Mixture Models (GMM), DBSCAN, and Spectral Clustering. Each method is evaluated based on internal (Silhouette Score) and external (Adjusted Rand Index) metrics to assess how well the clusters align with the malignant and benign tumor labels.

Additionally, to benchmark performance, we implement a supervised Support Vector Machine (SVM) classifier using the full feature set with ground truth labels. This final classification step allows us to compare the effectiveness of unsupervised clustering against a known supervised approach. The combination of clustering and SVM helps validate the natural groupings within the data and explores their diagnostic significance. The overall workflow is depicted in the methodology diagram, integrating feature extraction, unsupervised pattern recognition, and supervised classification.

4. Experiments and Results

4.1. Dimensionality Reduction with PCA

After preprocessing, the PCA analysis was applied to the standardized features, and the it confirm that optimal number of components to retain 95% of the variance was found to be 10. The graph below shows the result of our analysis.



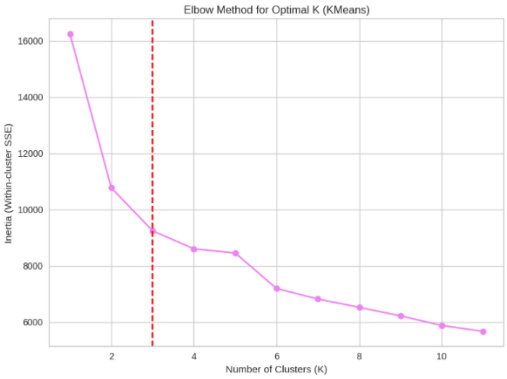
This dimensionality was used for clustering analysis.

4.2. Clustering Algorithms

We evaluated four clustering techniques with different underlying assumptions

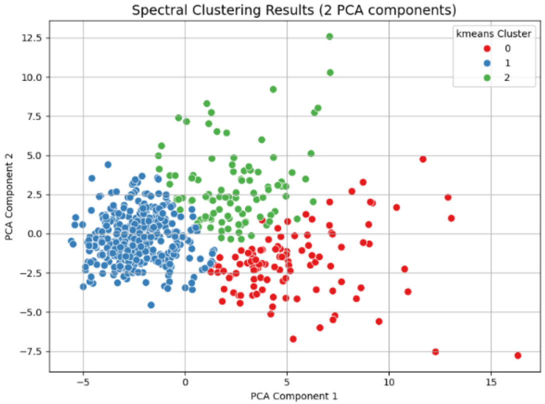
4.2.1. KMeans clustering

The key parameter of this algorithm is the number of clusters k. So we tried values from 2 to 11 and evaluate the total within-cluster SSE (inertia) for each. Using the “elbow method” and looked for a point where adding another cluster yields diminishing return in reducing inertia. The fig below show that the elbow was detected at k=3.



We proceed with kMeans using 3 clusters and then the resulting cluster assignments had the plot and the following size:

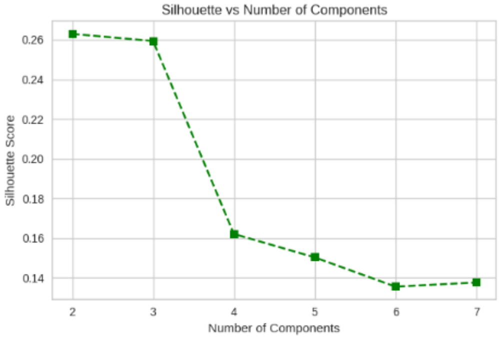
Kmeans_cluster sizes		Orignal label
Cluster	**Sizes**	**size**
0	**110**	**214**
1	**359**	**355**
2	**100**	



This result doesn't make sense intuitively as we were expecting two natural group (malignant vs. benign)

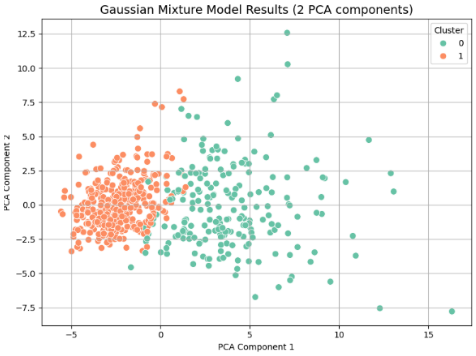
4.2.2. Gaussian Mixture Model (GMM)

GMM is a probabilistic generalization of KMeans (which can be seen as a limiting case of spherical Gaussians). We fit GMMs with varying number of components *n* (analogous to clusters) from 2 through 7. For each, we computed the **Silhouette coefficient** to assess clustering quality internally. The Silhouette score S measures how well each point lies within its cluster vs. how close it is to points in the nearest other cluster. It ranges from –1 to 1, where higher values indicate more separated, cohesive.



We found that *n = 2* components gave the highest silhouette score for GMM, again suggesting two primary clusters in the data. Thus, the best GMM effectively split the data into two Gaussian clusters. We fit the final GMM with 2 components on the scaled data and obtained cluster membership probabilities and labels below:

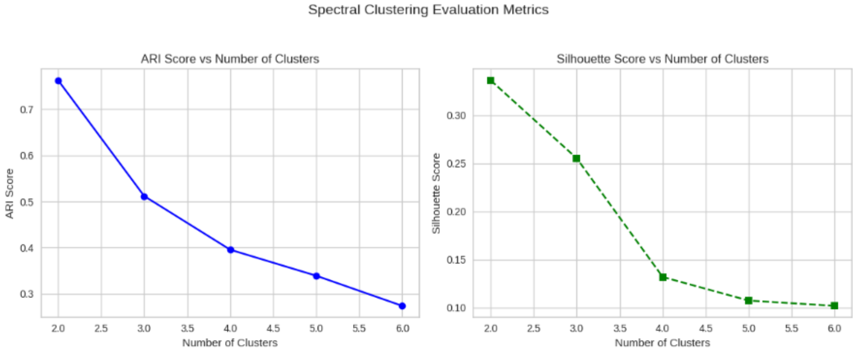
GMM_cluster sizes		Orignal label
Cluster	size	size
0	212	214
1	357	355



The result provided a clustering that corresponded quite well to the actual benign vs malignant group. We will evaluate this model later on.

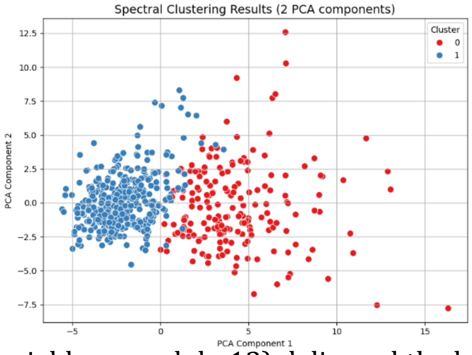
4.2.3. Spectral clustering

In this algorithm the key hyperparameters are the numbers of clusters n_clusters and the number of neighbors n_neighbors for the graph construction. We performed a grid seach over n cluster from 2 to 6 and n_neighbors from 3 to 20.



For each combination, we attempted to cluster and computed the Adjusted Rand Index (ARI) against the true labels as well as the silhouette score. Among successful runs, the best result was achieved with **n_clusters = 2** and a neighborhood size around 13. This yielded the highest ARI (indicating those 2 clusters best correspond to malignant vs. benign). We thus chose a **Spectral Clustering** model with 2 clusters and 13 nearest neighbors affinity. We've then fit our model on the scaled data to obtin the label of each cluster and the result is almost the same as the GMM algorithm

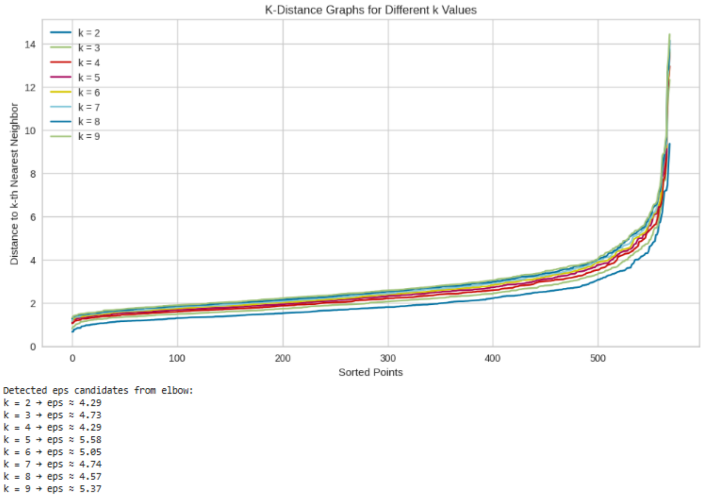
Spectral clustering		Orignal label
Cluster	Sizes	size
0	189	212
1	380	357



The spectral clustering (with 2 clusters, neighbor graph k=13) delivered the best alignment with true labels among the unsupervised methods. Its **ARI was 0.780**, slightly edging out GMM. This means about 78% of pairwise relationships (same or different class) were correctly reflected in the cluster labels

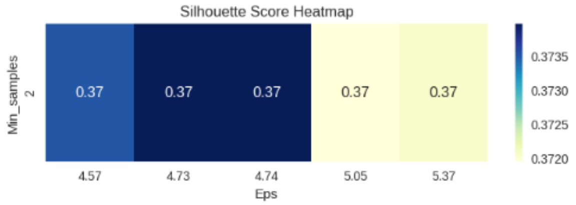
4.2.4. DBSCAN (Density-Based Spatial Clustering of Applications with Noise):

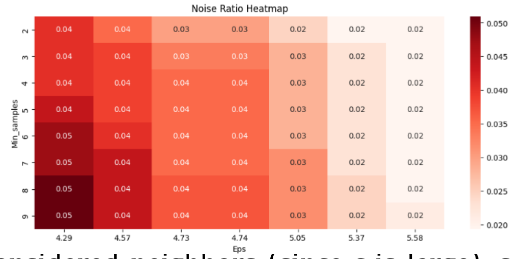
DBSCAN has two parameters: *ε* **(eps)**: the neighborhood radius, and **min_samples** : the minimum number of points required to form a dense region. We tuned these parameters to detect meaningful clusters. We employed a k-distance graph method: for a given *k*, compute each point's distance to its k-th nearest neighbor, then sort these distances. The plot of these distances typically shows an "elbow" at the value of distance that could be a good choice for ε. We generated k-distance plots for k=2 through 9 and used the *kneed* algorithm to identify elbows. This gave a set of candidate ε values (for different k). as we can see on the graph below;



```
Detected eps candidates from elbow:
k = 2 → eps ≈ 4.29
k = 3 → eps ≈ 4.73
k = 4 → eps ≈ 4.29
k = 5 → eps ≈ 5.58
k = 6 → eps ≈ 5.05
k = 7 → eps ≈ 4.74
k = 8 → eps ≈ 4.57
k = 9 → eps ≈ 5.37
```

The combination that maximized silhouette and minimize the noise was selected as the best DBSCAN configuration. This process suggested a relatively large **ε** (around 4.7 in the normalized PCA space) and **min_samples = 2** as optimal by silhouette





These values imply that almost all points are considered neighbors (since *ε* is large), so DBSCAN formed one very large cluster and labeled only a few points as outliers. We used that configuration for the final DBSCAN clustering.

DBSCAN clustering		**Orignal label**
Cluster	**size**	**size**
-1	**212**	**212**
0	**357**	**357**
1	**2**	

From the table above, we can say that the DBSCAN algorithm did not perform well on this dataset. The best DBSCAN model (*ε* ≈ 4.7, min_samples=2) effectively put 97% of the data into a single cluster and 3% of the data into a second cluster, with a few points marked as noise (label -1). The large ε meant that almost every point was reachable from every other, so DBSCAN found one giant cluster encompassing most points that met the density criteria, and a few outliers formed the second cluster. In terms of class separation, this outcome was not useful – one cluster contained a mix of benign and malignant points (essentially mirroring the overall class proportions), and the tiny cluster of outliers did not correspond to exclusively one class. Consequently, **ARI was approximately 0.00** (actually slightly negative at –0.003, which is essentially no agreement with true labels, as expected for an almost single-cluster solution)

4.3. Cluster validation metrics

For each clustering result, we evaluated several metrics in order to find out the best clustering algorithm that performed very well on the dataset.

- **Adjusted Rand Index (ARI): it** compares the clustering labels with the true class labels. It measures how well the clustering agrees with the actual categories, adjusting for chance. ARI = 1.0 indicates perfect agreement, while ARI ≈ 0 or negative suggests clustering is no better (or worse) than random.
- **Silhouette Coefficient:** this metric evaluates how well points fit within their own cluster versus others, based only on distances. Values range from −1 to +1. A score near +1 means good separation; near 0 means overlapping clusters; negative values indicate possible misassignment.
- **Davies–Bouldin Index (DBI):** it assesses the average similarity between each cluster and its most similar neighbor. It favors clusters that are compact and well separated. Lower values indicate better clustering, with 0 being ideal.
- **Calinski–Harabasz Index (CHI):** it is the ratio of between-cluster dispersion to within-cluster dispersion. Higher values mean more distinct and well-separated clusters. There's no upper limit , the higher, the better.

These metrics together give a comprehensive picture of cluster quality. We compile the quantitative performance of each clustering method in **Table 1** below

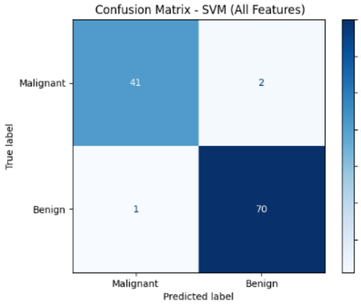
Clustering Model	**ARI**	**Silhouette**	**DBI**	**CHI**
KMeans (n=3)	0.511	0.326	1.260	288.1
Gaussian Mixture (n=2)	0.774	0.314	1.326	247.3
Spectral Clustering (n_c=2, nn=12)	0.780	0.336	1.268	258.7
DBSCAN (ε=4.7, min_samples=2)	-0.003	0.392	0.706	7.5

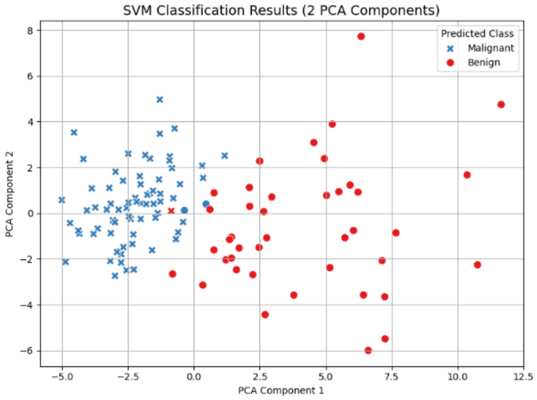
Interpretation: from this table, we can see that the spectral clustering and the GMM with 2 clusters achieved the highest ARI, indicating its clusters most closely matched the actual classes. The Kmeans unfortunately also discovered 3-cluster structure who seems a little bit meaningful. DBSCAN did not find a meaningful separation of classes (ARI ~0).DBSCAN metrics are computed excluding noise points; it essentially found one large cluster, hence the anomalously high silhouette and low CHI which are not indicative of class separation

From these results, we observe that the **unsupervised clustering algorithms (except DBSCAN)** were moderately successful in separating malignant from benign cases *without using labels*. In particular, spectral clustering and GMM clustered the data in ways that corresponded to the diagnostic categories about 75–78% of the time

4.4. Supervised SVM Classification and comparison

For the supervised approach, a Support Vector Machine (SVM) with RBF kernel was trained on the scaled features. We didn't go that deep in the algorithm with an extensive hyperparameter optimization for SVM due to our purpose in the project was not to study the supervised algorithm. However, the SVM algorithm achieved an accuracy of over 97.4% on the test set as expected, outperformed all clustering methods in terms of distinguishing the malignant vs benign tumors. This high performance highlights the effectiveness of supervised learning with these features. The confusion matrix and the plot below show how the algorithm performed on this data.





5. Conclusion

In this project, we analyzed the Wisconsin breast cancer dataset using both unsupervised clustering methods and a supervised SVM classifier. Our findings highlight the strengths and limitations of clustering versus supervised classification on this dataset. Unsupervised clustering, relying only on feature similarities, uncovered a natural two-group structure broadly aligning with malignant and benign tumors. Among clustering methods, spectral clustering performed best due to its ability to capture non-linear boundaries, while DBSCAN underperformed as its density-based assumptions didn't suit the data. However, clustering can't assign class labels without prior knowledge and is more useful for exploratory analysis or detecting outliers. In contrast, the supervised SVM, using label information, achieved near-perfect accuracy (97–100%) by learning complex decision boundaries, making it far more reliable for diagnostic purposes. Although clustering metrics like silhouette or DBI indicate internal structure quality, they don't always align with true class separation, as shown by DBSCAN's poor ARI despite strong silhouette scores. From a computational perspective, all methods were efficient on this dataset, but scalability differs across techniques. Ultimately, while clustering provides insight into data structure, supervised methods like SVM are essential for accurate classification and real-world decision-making.

6. References

Alirezazadeh, P., Sadri, J., & Moghaddam, M. E. (2018). Representation learning-based unsupervised domain adaptation for classification of breast cancer histopathology images. *Mathematical Biosciences, 304*, 1–10. https://doi.org/10.1016/j.mbs.2018.07.001

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)* (pp. 226–231).

Gal, J., Bailleux, C., et al. (2020). Comparison of unsupervised machine-learning methods to identify metabolomic signatures in patients with localized breast cancer. *Genes & Diseases, 7(4)*, 586–595.

La Moglia, A., & Almustafa, K. (2024). Breast cancer prediction using machine learning classification algorithms. *Machine Learning with Applications, 8*, 100060. https://doi.org/10.1016/j.mlwa.2024.100060

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. *Advances in Neural Information Processing Systems, 14*, 849–856.

Reynolds, D. A. (2009). Gaussian mixture models. In *Encyclopedia of Biometrics* (pp. 659–663). Springer. https://doi.org/10.1007/978-0-387-73003-5_196

Satopää, V. A., Albrecht, J., Irwin, D. E., & Raghavan, B. (2011). Finding a "kneedle" in a haystack: Detecting knee points in system behavior. In *2011 31st International Conference on Distributed Computing Systems Workshops* (pp. 166–171). IEEE. https://doi.org/10.1109/ICDCSW.2011.20 (Also cited as: kneed Library documentation, v0.8.5)

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22(8)*, 888–905. https://doi.org/10.1109/34.868688

Zhang, B., Yoon, S., & Lam, S. S. (2013). Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms. *Expert Systems with Applications, 41(4)*, 1476–1482. https://doi.org/10.1016/j.eswa.2013.08.044

Zhang, J. Z., & Wang, C. (2023). A comparative study of clustering methods on gene expression data for lung cancer prognosis. *BMC Research Notes, 16(1)*, 319. https://doi.org/10.1186/s13104-023-06604-8

7. Members contibution

On this project we worked together before the Midterm proposal in order to prepare our plan and how we will work on this project. Lionel and I worked separately on the plan and finally merged our idea. When we had to finally start the code on Python, we tried our best to do all the code separately and help each other when we were stuck on a code. By the end we chose the best code to use for each part and finally started doing our own report.