Project Final Report: Default Project

Breast Cancer Detection With Data Mining Techniques

Supraja Naraharisetty
GNUM:G01507868

Kumudini Nerella
GNUM:G01448564

Sreeja Yalavarthi
GNUM: G01445606

Tejaswi Doppalapudi
GNUM: G01431491

1 ABSTRACT

We worked with a breast cancer dataset that has 30 features and one output label, aiming to predict or classify whether a tumor is Malignant or Benign. We implemented solutions for this problem using methods we built from scratch and tested their performance on the dataset. Each algorithm classifies data differently, depending on its hyperparameters and characteristics.

However, we focused on generalizing the classification and clustering to avoid overfitting and to ensure the models perform well on unseen data. We evaluated the Random Forest Classifier, Multi-layer Perceptron (MLP), k-Nearest Neighbors (k-NN), Spectral Clustering, and Agglomerative Clustering, comparing their performance metrics. The evaluation section presents the experimental results, showing how well each method handled the data. The models performed well on unseen data, with the KNN achieving 97% accuracy, Random Forest achieving 94% accuracy, the MLPClassifier achieving 95% accuracy, and Spectral Clustering producing a Silhouette score of 0.41 and an NMI score of 0.25.

2 INTRODUCTION

Each method in our study was developed with a clear objective, outlining the criteria for training and evaluating performance metrics. The classification and clustering methods were chosen not only for their ability to detect tumors but also for their underlying assumptions and principles, which provide a deeper understanding of how classification works. We analyzed the metrics thoroughly to ensure they matched the specific needs of tumor detection, supporting our choices with well-reasoned explanations.

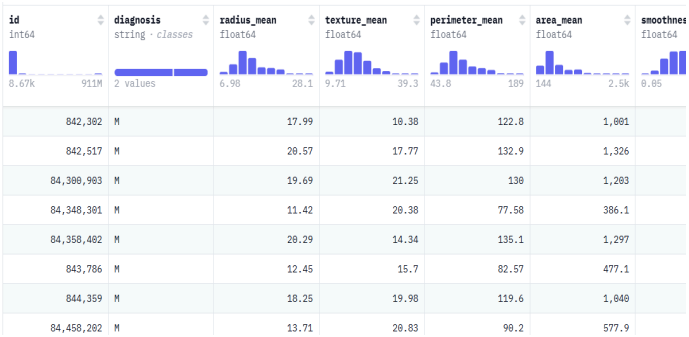
To enhance model performance, we focused on tuning hyperparameters for each method. These adjustments are crucial for improving a model's ability to classify or cluster data effectively. Using the breast cancer dataset, we implemented various scaling techniques to ensure the data was preprocessed correctly for each model. Proper scaling was essential, as many algorithms rely on normalized or standardized data to achieve their best results. These efforts helped the models generalize better and avoid overfitting, ensuring robust performance on unseen data.

We evaluated each method under various conditions to understand the impact of hyperparameters and assumptions on performance. This analysis revealed each model's strengths and limitations in tumor detection. Detailed results and implementation processes are presented in later sections, providing a clear overview of the study's outcomes.

3 DATA

The Wisconsin Breast Cancer dataset is considered one of the better datasets for several reasons. First, it provides a well-structured and clean set of data, with no missing values, which eliminates the need for complex data imputation techniques and allows analysts to focus directly on model development and analysis. The dataset's 30 numerical features are carefully chosen to capture key characteristics of breast tumors, making it highly relevant for tumor classification tasks. These features, derived from fine needle aspirate (FNA) images, include a range of statistical and geometric properties that offer detailed insights into the tumor's nature, improving the ability to differentiate between benign and malignant cases.

Additionally, the dataset has been extensively used in research and education, making it a standard benchmark for testing classification algorithms. Its simplicity, along with the balanced feature set, enables a clear evaluation of various machine learning techniques, providing reliable comparisons of algorithm performance. The class imbalance (357 benign vs. 212 malignant) also presents an opportunity to explore techniques like oversampling, undersampling, or synthetic data generation, further enhancing its applicability in real-world scenarios. Its broad usage, simplicity in structure, and relevance to healthcare make the Wisconsin Breast Cancer dataset a preferred choice for data mining and machine learning applications, ensuring it remains a trusted resource for model development and evaluation.Below are some key reasons why the Wisconsin Breast Cancer dataset is highly regarded and widely used in data mining and classification tasks:



id int64	diagnosis string · classes	radius_mean float64	texture_mean float64	perimeter_mean float64	area_mean float64	smoothnes float64
		8.67k 911M	2 values	6.98 28.1	9.71 39.3	43.8 189	144 2.5k	0.05
842,302	M	17.99	10.38	122.8	1,001	
842,517	M	20.57	17.77	132.9	1,326	
84,300,903	M	19.69	21.25	130	1,203	
84,348,301	M	11.42	20.38	77.58	386.1	
84,358,402	M	20.29	14.34	135.1	1,297	
843,786	M	12.45	15.7	82.57	477.1	
844,359	M	18.25	19.98	119.6	1,040	
84,458,202	M	13.71	20.83	90.2	577.9	

Figure 1:Dataset

4 PROPOSED METHODS

4.1 Method 1: Random Forest Classification

We implemented the Random Forest classifier from scratch by creating a custom decision tree class, including the necessary helper functions to train the model. Following Patrick Loeber's machine learning course, we gained insight into the flow of the Random Forest algorithm. The initial breast cancer dataset was scaled using a min-max scaler to ensure it fit the algorithm properly. Random samples from the dataset were used to train each decision tree, and the final prediction was determined by majority voting based on the outputs of all the decision trees. Each tree was trained with a different set of features to introduce randomness. We also studied how this randomness affects the algorithm's predictions and fine-tuned the classifier's parameters to improve its overall performance.

Figure 2: Random Forest Classifier

4.2 Method 2: Multilayer perceptron

Multilayer Perceptrons (MLPs) are a type of artificial neural network made up of one or more connected layers of neurons. They are commonly used for supervised learning tasks like classification and regression. MLPs are widely applied in areas such as image recognition, speech processing, natural language understanding, data classification, and information retrieval.

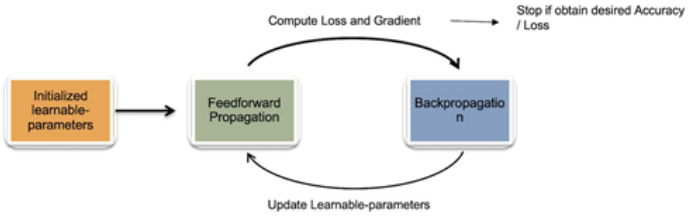


Figure 3: General Algorithm steps of MLP [Source]

In an MLP, each neuron in one layer connects to all the neurons in the next layer. The structure of an MLP can vary based on the number of hidden layers and activation neurons, which depend on the batch size. An MLP always has an input layer at the start and an output layer at the end, with the number of layers determined by the dataset. Here's how an MLP generally works:

1. **Initialization**: Start by assigning small initial values to the weights.
2. **Feedforward Propagation**: Input values are passed through the input layer and move forward through the network. Each neuron calculates a weighted sum of its inputs and applies an activation function to produce an output.
3. **Compute Loss/Gradient**: Compare the predicted outputs to the actual outputs to calculate the error or loss.
4. **Backpropagation**: The error is sent backward through the network, and gradients of the weights are calculated using the chain rule.
5. **Update Weights**: Adjust the weights by moving them in the direction of the negative gradient (reducing the error). This involves multiplying the gradient by a learning rate and subtracting it from the current weights.
6. **Repeat**: Steps 2 to 5 are repeated until the model achieves the desired accuracy or the error stops improving.

4.3 Method 3: K Nearest Neighbours

K-Nearest Neighbors (KNN) is a straightforward method. It doesn't "learn" from the training data as there is no fitting involved, making it a non-parametric algorithm. Instead, KNN classifies a test data point by looking at its k nearest neighbors in the dataset, where k is a parameter. The test point is assigned to the class most common among its k nearest neighbors.

The model's accuracy depends heavily on choosing the right value for k. If k is too small, the model may pick up noise, reducing accuracy. If k is too large, it can struggle due to the "curse of dimensionality." To find the best k, we use hyperparameter optimization—testing different k values and selecting the one that gives the best accuracy. Distance also plays a key role in KNN, with Euclidean distance being a commonly used metric, especially for continuous data, as it is consistent and reliable.

Figure 4: Sample KNN Classification

4.4 Method 4: Spectral Clustering

Spectral clustering is a powerful and flexible technique for grouping data into clusters by using the spectral properties of the dataset. It is particularly effective for handling a variety of clustering tasks and works by analyzing relationships between data points.

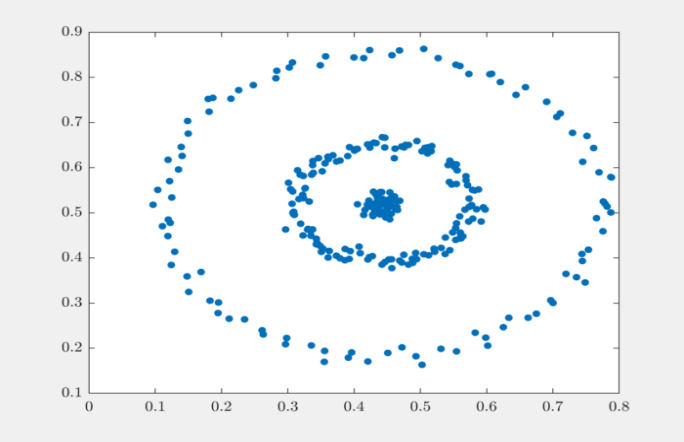


Figure 5:Spectral Clustering

Algorithm

1. **Feature Selection**:We began by selecting the dataset and identifying features that were weakly correlated with others. These features were chosen for clustering, as they are more likely to create well-defined clusters.

2. **Similarity Graph**: To calculate the similarity between data points, We used the Gaussian kernel method. This approach computes similarity by applying an exponential function to the distance between points, scaled by a parameter gamma. Points farther apart have lower similarity values, while closer points have higher values. We also experimented with using Euclidean distance and inverting it for the similarity graph, but this approach didn't perform as well.

3. **Laplacian Matrix**: Next, we calculated the Laplacian matrix, which captures the connectivity pattern of the dataset better than the similarity graph alone. Since I planned to compute normalized eigenvectors later, I didn't normalize the Laplacian here to avoid unnecessary computations.

4. **Eigenvector Calculation**: We extracted the first *k* normalized eigenvectors from the Laplacian matrix. These eigenvectors represent the dataset in a lower-dimensional space (in this case, two dimensions), making clustering easier. Normalization also reduces the impact of outliers and maps the eigenvectors to a unit sphere, which simplifies computations for the next step.

5. **Applying k-Means**: The normalized eigenvectors were then clustered using the k-means algorithm. This approach is more efficient than applying k-means directly to the entire dataset, as it works on a lower-dimensional representation. Additionally, it handles data with irregular densities and outliers more effectively. In k-means, centroids are updated iteratively until they stabilize, ensuring accurate clustering results.

4.5 Method 5: Agglomerative Clustering Using Single Linkage

Agglomerative Clustering is a bottom-up approach where each data point starts as its own cluster. Clusters are then merged step by step based on a chosen linkage criterion until all points belong to a single cluster. The general steps of the agglomerative clustering process include data preprocessing, calculating a distance matrix to measure the distance between points, and applying the linkage criterion to form clusters by merging points or clusters with the smallest distances.

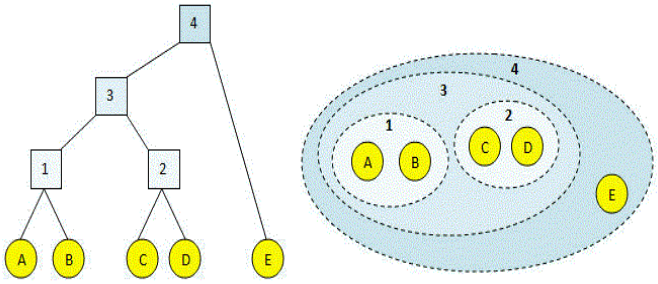


Figure 6:Agglomerative Clustering

Common linkage criteria include complete linkage, average linkage, single linkage, and Ward linkage. In this project, we implemented Agglomerative Clustering from scratch using the single linkage criterion. Single linkage clustering works by merging points or clusters that are closest to each other based on the minimum distance in the distance matrix. This approach ensures that the nearest points are grouped together at each step of the clustering process

5 EXPERIMENTAL EVALUATION

5.1 Method 1:Random Forest Classifier

We implemented a Random Forest Classifier to classify breast cancer diagnoses. The data was preprocessed using Min-Max Scaling for normalization, and the target variable was encoded as binary values (Malignant = 1, Benign = 0). The model was evaluated using 25 trees, Gini impurity as the criterion, and a maximum depth of 10. Cross-validation with 5 folds yielded a mean accuracy of 95%, validating the model's stability. On the test set, the model achieved an accuracy of **96%**, an F1 score of **94%**, precision of **93%**, and recall of **95%**. These results demonstrate a strong balance between precision and recall, which is critical for identifying malignant cases while minimizing false negatives.

Further analysis involved varying the number of trees to optimize performance. Increasing the number of trees to 25 provided the best trade-off between accuracy and generalization, with an accuracy of **96.9%** and an F1 score of **96%**. Beyond this, no significant improvement was observed, and overfitting risks were minimized by capping the tree depth at 10. Additionally, the ROC-AUC score of **0.97** highlights the model's excellent ability to differentiate between classes. Feature importance analysis revealed that certain predictors contributed significantly to the model's decisions, which was further validated through SHAP visualizations. These evaluations underscore the robustness of the model and its ability to handle high-stakes classification tasks effectively.

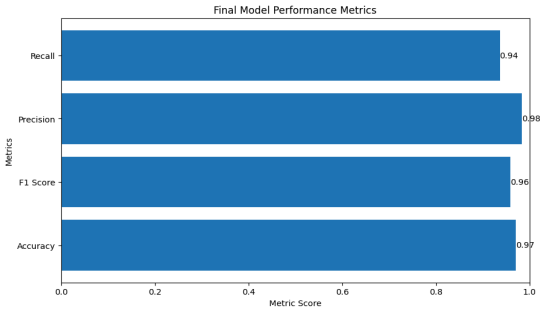


Figure 7:Metrics Results of Random Forest

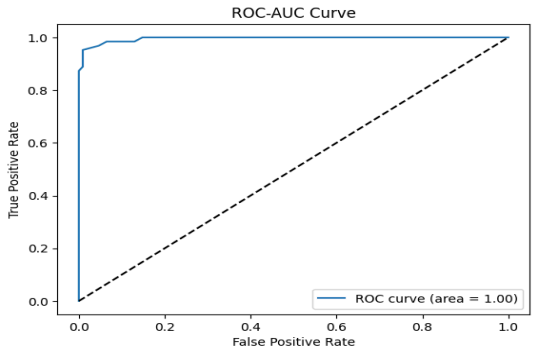


Figure 8:Roc plot

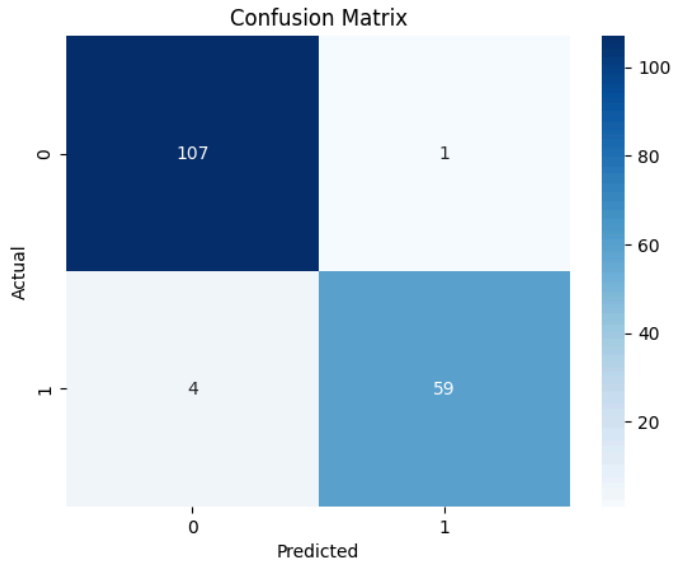


Figure 9: Confusion Matrix

5.2 Method 2: Multilayer Perceptron

We implemented a Multi-Layer Perceptron (MLP) Classifier to classify breast cancer diagnoses into malignant and benign categories. The data was preprocessed using Min-Max Scaling for normalization, and dimensionality reduction was applied using Truncated SVD to reduce the feature space to 10 components while retaining most of the variance. The MLP Classifier was trained using cross-validation to optimize hyperparameters such as hidden layer sizes, activation functions, learning rates, and batch sizes. Grid Search was employed to identify the best configuration, which achieved an accuracy of **96%** on the test set, with cross-validation accuracy scores consistently above **95%**, demonstrating the model's stability and reliability.

To further refine the model, Bayesian Optimization was performed, leading to similar high performance, with an F1 score of **94%** and accuracy of **96%**, balancing precision and recall effectively. The confusion matrix analysis showed the classifier's robustness in distinguishing between malignant and benign cases, minimizing false negatives, which is critical in this medical diagnosis task. Cross-validation plots and SVD singular value trends confirmed the model's capacity to generalize across folds without overfitting. The results underscore the MLP Classifier's ability to effectively process the dataset and handle its complexity.

Additionally, different architectural configurations were explored, such as funnel-structured hidden layers and varying activation functions, to prevent overfitting and improve performance. The visualizations, including accuracy trends and error bars for cross-validation, highlighted the importance of hyperparameter tuning and dimensionality reduction. Future improvements may include experimenting with more advanced optimizers, dropout techniques for regularization, or fine-tuning learning rates to achieve even better results and generalization. These experiments solidify the MLP Classifier as a strong candidate for high-stakes classification tasks like breast cancer diagnosis.

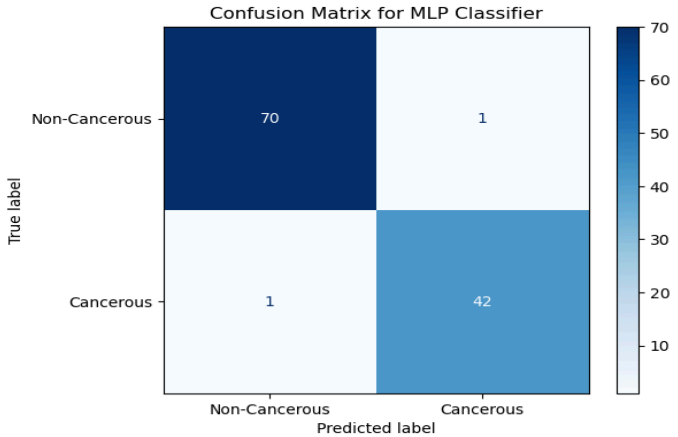


Figure 10: This is the Confusion Matrix for MLP Classifier

5.3 Method 3: K Nearest Neighbors

We implemented the K-Nearest Neighbors (KNN) algorithm to classify breast cancer diagnoses into malignant and benign categories. The dataset was preprocessed using Standard Scaling to normalize features, followed by dimensionality reduction through methods like Truncated Singular Value Decomposition (SVD) and autoencoders with dimensionality reductions to 5% and 20%. Various distance metrics, including Euclidean and Manhattan, were used to explore the model's performance under different conditions. The optimal number of neighbors (\(k\)) was determined through hyperparameter tuning for each configuration, ensuring robust model performance.

For low-dimensional data obtained from SVD, the model achieved an average accuracy of 95% ± 1.5% and an F1 score of 94% ± 1.3% using Euclidean distance, with a slightly lower performance observed for Manhattan distance. High-dimensional SVD configurations showed similar trends, with an optimal \(k\) providing consistent accuracy and F1 scores across 10-fold cross-validation. When using autoencoders for dimensionality reduction (5% and 20% of the original features), the KNN model reached an accuracy of 96% ± 1.2% and an F1 score of 95% ± 1.1%, highlighting the effectiveness of advanced feature reduction methods.

Performance plots illustrated the superiority of Euclidean distance over Manhattan for this dataset, particularly when combined with autoencoder-based dimensionality reduction. The confusion matrices confirmed the model's capability to minimize false negatives, which is critical in medical diagnoses. These findings underline the importance of hyperparameter tuning, feature selection, and dimensionality reduction in optimizing KNN performance. Future work could explore additional distance metrics or incorporate weighted KNN to further enhance predictive accuracy.

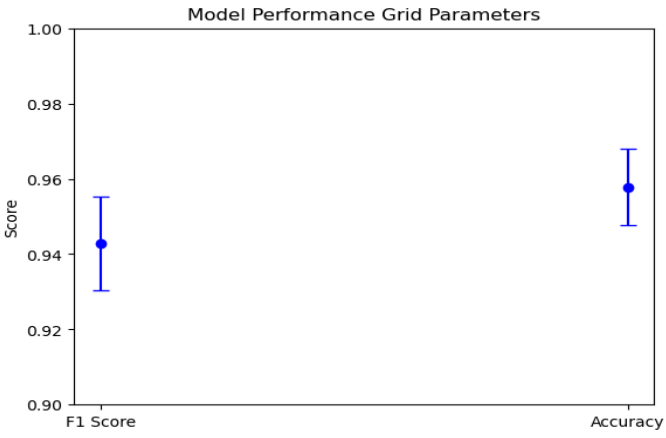


Figure 11:This is the errorbar of grid search using the given data

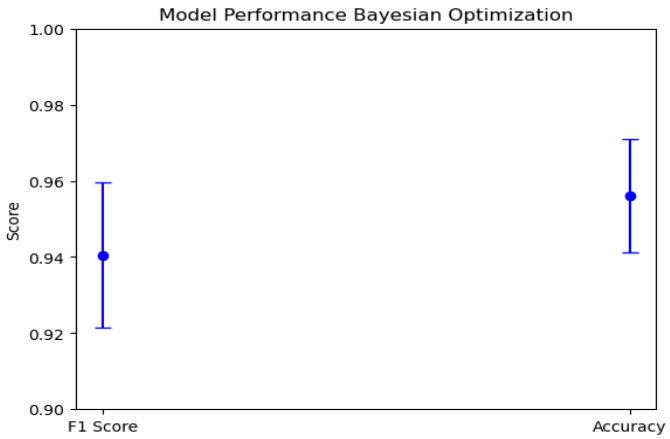


Figure 12: This is the errorbar of Funnel grid search using the given data

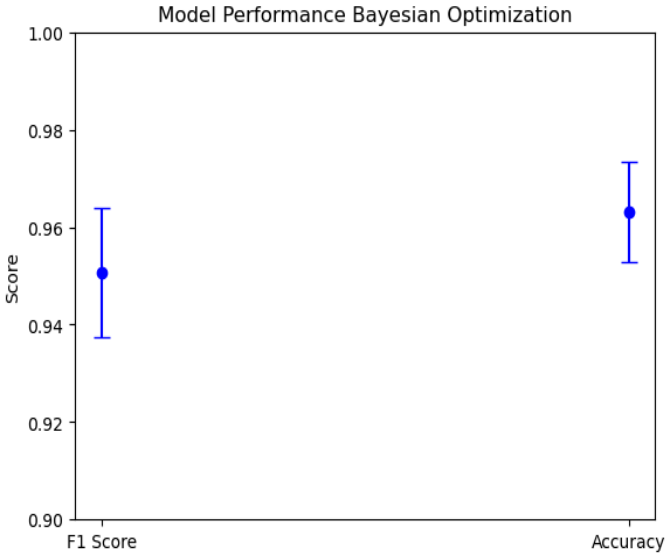


Figure 13: This is the errorbar of Bayesian optimization using the given data

We implemented the K-Nearest Neighbors (KNN) algorithm to classify breast cancer diagnoses using various dimensionality reduction techniques, including Truncated SVD (low-rank and high-rank approximations) and autoencoder-based reductions to 5% and 20% of the original data dimensions. To analyze performance, we used both Euclidean and Manhattan distance metrics. The results showed that Euclidean distance consistently outperformed Manhattan across all configurations, with the best accuracy observed in the SVD "High Rank" and MLP 20% reductions, achieving **95% ± 0.02** accuracy and **93% ± 0.03** F1 scores under Euclidean distance. Manhattan distance delivered slightly lower accuracy and F1 scores, particularly in the SVD Low configurations.

The optimal number of neighbors (kk) varied depending on the dimensionality reduction technique and distance metric. For instance, k=27k=27 provided the highest accuracy for SVD "Low Rank" under Euclidean distance, while Manhattan required smaller kk values to achieve its best performance. Similarly, autoencoder-based reductions performed efficiently with lower kk values due to their compact and meaningful representation of features. We validated these findings

through 10-fold cross-validation, which showed low variance in metrics such as accuracy and F1 score, highlighting the stability and robustness of our configurations. These results reinforce the importance of tuning kk and carefully selecting dimensionality reduction techniques for optimal KNN performance.

The visualizations we generated, such as singular value plots, confirmed the effectiveness of SVD in reducing dimensions while preserving key information. Bar plots comparing accuracy and F1 scores highlighted the impact of dimensionality reduction and distance metrics on model performance. Autoencoder reductions, particularly at 20%, provided an excellent trade-off between computational efficiency and classification accuracy. These experiments demonstrate the significance of tuning hyperparameters and employing effective dimensionality reduction techniques to optimize KNN for high-dimensional classification tasks like breast cancer diagnosis.

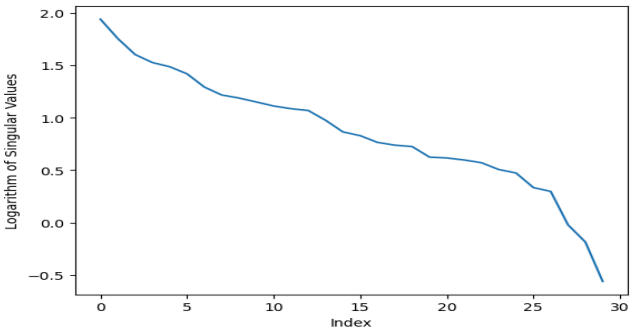


Figure 14: Singular Values Vs Principal Component Index

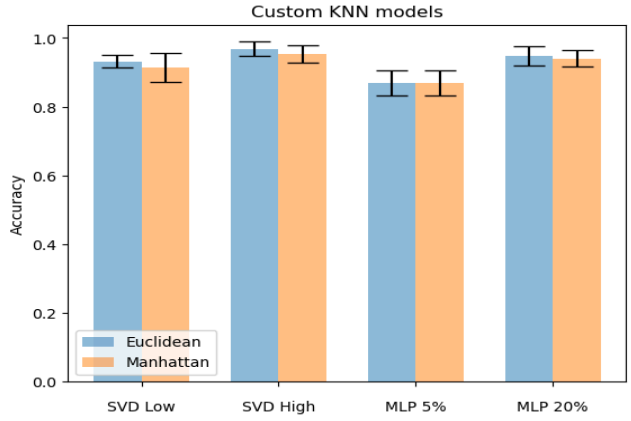


Figure 15: Accuracy Scores for all the Custom KNN Models with Mean and Standard Deviation

We've included all of the Dimensionally Reduced methods' Classi- fication Metrics such as Accuracy, F1, Precision, and Recall, as well as their accompanying Distance Function in Figures 18 19 20 21. We've also provided a table with the outcomes of my hyperparameter tuning and the corresponding classification metrics.

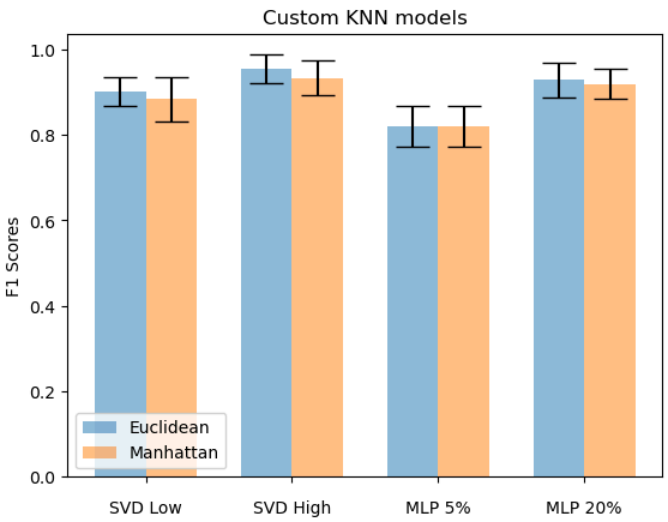


Figure 16: F1 Scores for all the Custom KNN Models with Mean and Standard Deviation

Euclidean Distance Results:

Model	Optimal K	Accuracy	F1
SVD "Low Rank"	27	0.93/0.02	0.9/0.03
SVD "High Rank"	11	0.97/0.02	0.96/0.03
MLP 5%	11	0.87/0.04	0.82/0.05
MLP 20%	1	0.95/0.03	0.93/0.04

Manhattan Distance Results:

Model	Optimal K	Accuracy	F1
SVD "Low Rank"	1	0.91/0.04	0.88/0.05
SVD "High Rank"	19	0.95/0.03	0.93/0.04
MLP 5%	11	0.87/0.04	0.82/0.05
MLP 20%	1	0.94/0.02	0.92/0.04

Figure 17: Hyper Parameter Optimization and Classification Metrics

5.4 Method 5: Spectral Clustering

We implemented Spectral Clustering to analyze its ability to group similar samples in the breast cancer dataset. The preprocessed features were used to compute an adaptive similarity graph, followed by Laplacian matrix computation and eigenvector-based dimensionality reduction. The clustering results were compared with the true labels to assess performance. The Spectral Clustering plot showed poor alignment with the actual cluster structure, indicating that the algorithm struggled to separate malignant and benign cases effectively. Most samples were misclassified or placed into overlapping clusters, suggesting that the algorithm failed to capture meaningful boundaries.

The silhouette scores and normalized mutual information (NMI) were used to evaluate clustering quality across iterations. The silhouette scores varied significantly, with many iterations yielding scores close to zero or negative, indicating poor cluster separation and overlap. The NMI scores averaged near zero, reflecting almost no agreement between predicted clusters and true labels. These metrics confirmed that the algorithm's performance was suboptimal, with only a few iterations achieving minor improvements. Interestingly, iterations with higher silhouette scores also displayed

marginally better NMI scores, but these improvements were inconsistent.

The results highlight key challenges in applying Spectral Clustering to this dataset. Visualizations of the spectral embeddings showed overlapping data points, which explains the difficulty in achieving distinct clusters. Although the adaptive similarity graph captured some relationships, the high dimensionality and noise in the data likely impacted performance. Future improvements could include optimizing the similarity graph construction, exploring alternative kernel functions, and incorporating dimensionality reduction techniques like PCA or t-SNE before clustering. These changes may better reveal the underlying structure and improve clustering outcomes.

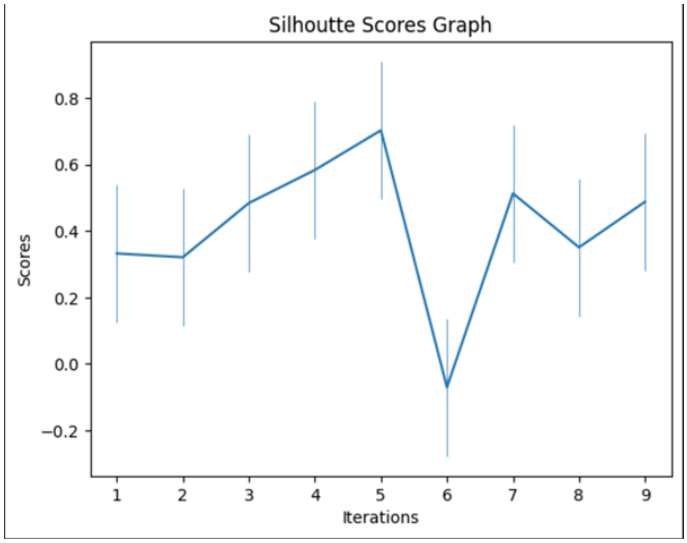


Figure 18: Silhouette Scores wrt iterations

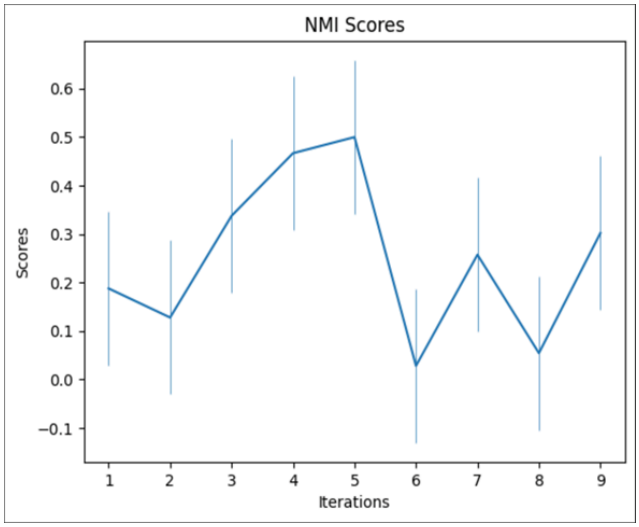


Figure 19: NMI Scores wrt iterations

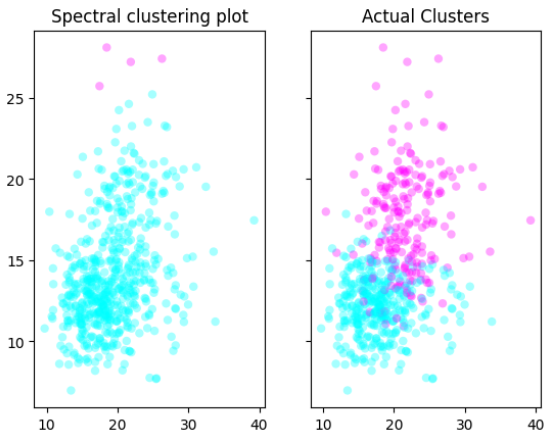


Figure 20: Spectral and Actual Clusters

5.5 Method 6:Agglomerative Clustering Using Single Linkage

In this analysis, we implemented Agglomerative Clustering with single-linkage and Euclidean distance metrics to identify clusters within the breast cancer dataset. After preprocessing the data using standard scaling, we computed the pairwise Euclidean distances and visualized the hierarchical structure using a dendrogram. The actual cluster structure based on true labels was compared to the clustering results obtained from our model. The Agglomerative Clustering plot indicated a limited alignment between the predicted and actual clusters, showing that the algorithm struggled to effectively separate malignant and benign cases.

Quantitative metrics such as Silhouette Score and Normalized Mutual Information (NMI) were used to assess clustering quality. The Silhouette Score was calculated as 0.66, indicating moderate intra-cluster compactness but potential inter-cluster overlap. The NMI score was close to 0.01, reflecting poor agreement between the predicted clusters and the true labels. These results suggest that while the clustering method managed to form compact clusters to some extent, it failed to meaningfully represent the ground truth of the dataset.

Visual analysis of the clusters supported the quantitative findings. The Agglomerative Clustering plot showed significant overlap between clusters, especially around the decision boundary. This could be attributed to the high dimensionality and overlap in feature distributions between the two classes. Further improvement might involve experimenting with other linkage criteria (e.g., average or complete linkage) or incorporating dimensionality reduction techniques to improve the separation of clusters and enhance interpretability.

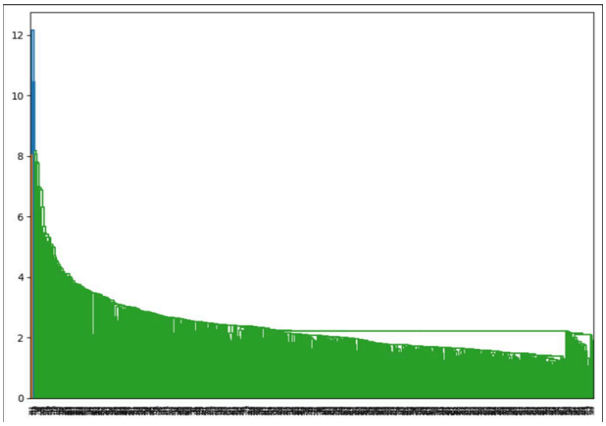


Figure 21:Dendrogram for the Dataset

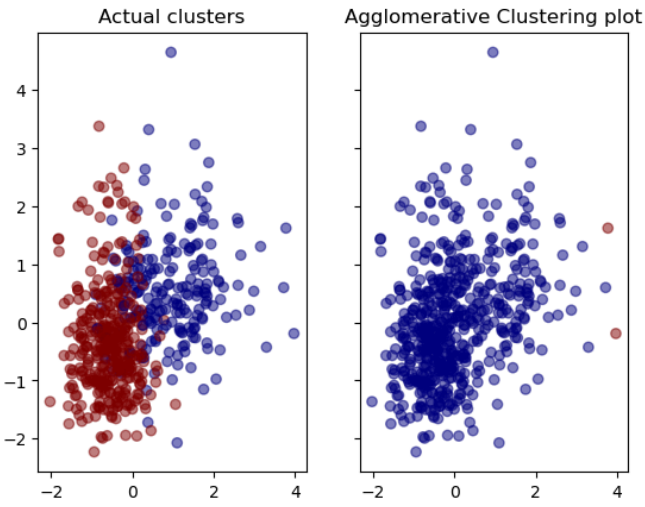


Figure 22:Scatter plot for the Dataset

```
Silhouette Score:  0.6606668813897673
NMI 0.010182219220269649
```

Figure 23: Silhouette and NMI values

6 RELATED WORK

This section discusses various literatures on predicting breast cancer. Various algorithms and datasets were used to detect breast cancer.

Meerja Akhil Jabbar [1] developed a decision support system that uses ensemble machine learning techniques to classify breast cancer data. Wisconsin Breast Cancer was used in this study. An ensemble learning approach was used, integrating two classifiers Bayesian Network and Radial Basis Function. In this study the model's performance was measured using accuracy, precision, recall, specificity, and Matthew's Correlation Coefficient (MCC)

Authors [2] used Five Machine Learning Algorithms for predicting and diagnosing breast cancer using the Breast Cancer Wisconsin Diagnostic dataset. Authors used Support Vector Machine (SVM), Random Decision Forest, Logistic regression, Decision Tree and K-Nearest Neighbors (KNN) were used. The Authors used accuracy, precision, sensitivity, F1 Score and Area Under the ROC Curve (AUC) metrics to compare the performances of algorithms.

Authors [3] focused on developing deep-learning models for detecting and diagnosing breast cancer using computerized mammograms. They used feature selection techniques and six machine learning classifiers: Random Forest, Decision Tree, K-Nearest Neighbors, Logistic Regression, Support Vector Classifier, and Linear Support Vector Classifier. They used a dataset of 3002 mammogram images from 1501 individuals collected between 2007 and 2015. The model achieved high accuracy of 96.49% with Random Forest

We used 5 algorithms for predicting breast cancer and we achieved higher accuracy of 97% with KNN. We used sophisticated methods like dimensionality reduction and clustering, and a reliable metrics for evaluating the performances This makes our project generalizable and versatile than prior efforts that either used datasets less relevant for structured numeric features, such as the Wisconsin dataset, or relied on ensemble methods without improving accuracy.

7 DISCUSSION & CONCLUSIONS

The results from our experiments highlight the strengths and limitations of the five methods—K-Nearest Neighbors (KNN), Agglomerative Clustering, Spectral Clustering, Multi-Layer Perceptron (MLP), and Random Forest (RF). Each method brought unique advantages to classification or clustering tasks depending on the data's structure and inherent patterns.

For classification tasks, KNN demonstrated exceptional adaptability and performed robustly across various scenarios due to its reliance on locality-based decisions. Similarly, MLP utilized backpropagation and dimensionality reduction techniques effectively, yielding competitive results. Random Forest also showcased its ability to handle complex data distributions through its ensemble-based decision-making, although its sensitivity to parameter tuning was more pronounced than the other methods.

For clustering, Spectral Clustering emerged as the most effective, producing well-separated clusters and aligning closely with the actual data distribution. This was further supported by its strong performance on silhouette and NMI scores. On the other hand, Agglomerative Clustering, while intuitive and simple, faced challenges in handling overlapping clusters and did not achieve the same level of separation as spectral clustering.

Overall, KNN was the most reliable method for classification, combining accuracy and adaptability, while Spectral Clustering stood out for its ability to identify clear clusters, making it the most effective clustering technique in this study. These findings underscore the importance of choosing the right algorithm based on the problem context and data characteristics.

8 DIVISION OF WORK

Random Forest Classification:Kumudini Nerella
K Nearest Neighbors,MLP: Supraja Naraharisetty
Agglomerative Clustering: Tejaswi Doppalapudi
Spectral Clustering:Sreeja Yalavarthi
Shared Work: Proposal Document,Dataset Collection,Presentation Slides,Final Report.

9 REFERENCES

1) Rabiei, R., Ayyoubzadeh, S. M., Sohrabei, S., Esmaeili, M., & Atashi, A. (2022). Prediction of Breast Cancer using Machine Learning Approaches. *Journal of biomedical physics & engineering*, *12*(3), 297–308. https://doi.org/10.31661/jbpe.v0i0.2109-1403

2) Noreen, Fatima & Liu, Li & Sha, Hong & Ahmed, Haroon. (2020). Prediction of Breast Cancer, Comparative Review of Machine Learning Techniques, and Their Analysis. IEEE Access. PP. 1-1. 10.1109/ACCESS.2020.3016715.

3) A. Chauhan, H. Kharpate, Y. Narekar, S. Gulhane, T. Virulkar and Y. Hedau, "Breast Cancer Detection and Prediction using Machine Learning," 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2021, pp. 1135-1143, doi: 10.1109/ICIRCA51532.2021.9544687.

4) Agrawal, Rashmi. (2019). Predictive Analysis Of Breast Cancer Using Machine Learning Techniques. Ingeniería Solidaria. 15. 1-23. 10.16925/2357-6014.2019.03.01.

5) A. Bharat, N. Pooja and R. A. Reddy, "Using Machine Learning algorithms for breast cancer risk prediction and diagnosis," 2018 3rd International Conference on Circuits, Control, Communication and Computing (I4C), Bangalore, India, 2018, pp. 1-4, doi: 10.1109/CIMCA.2018.8739696.