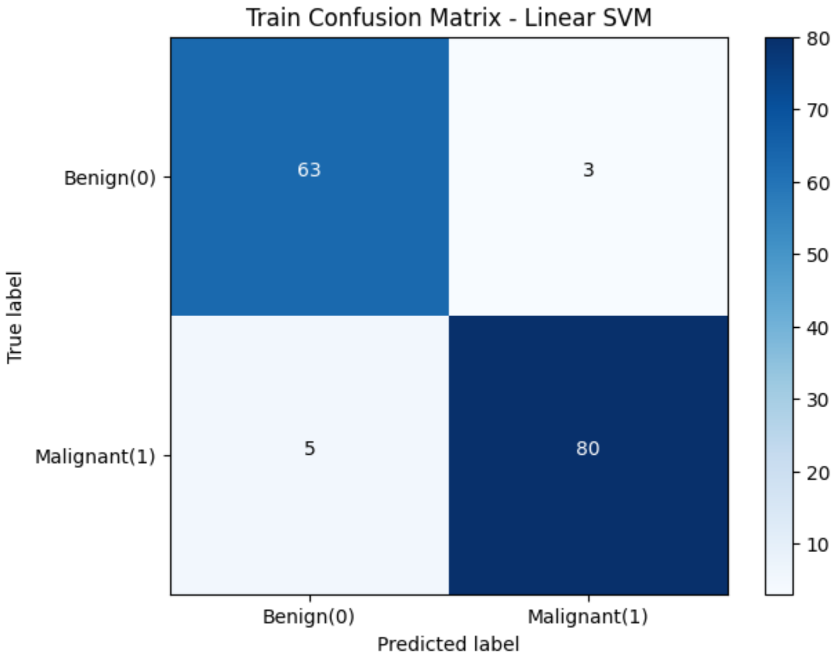
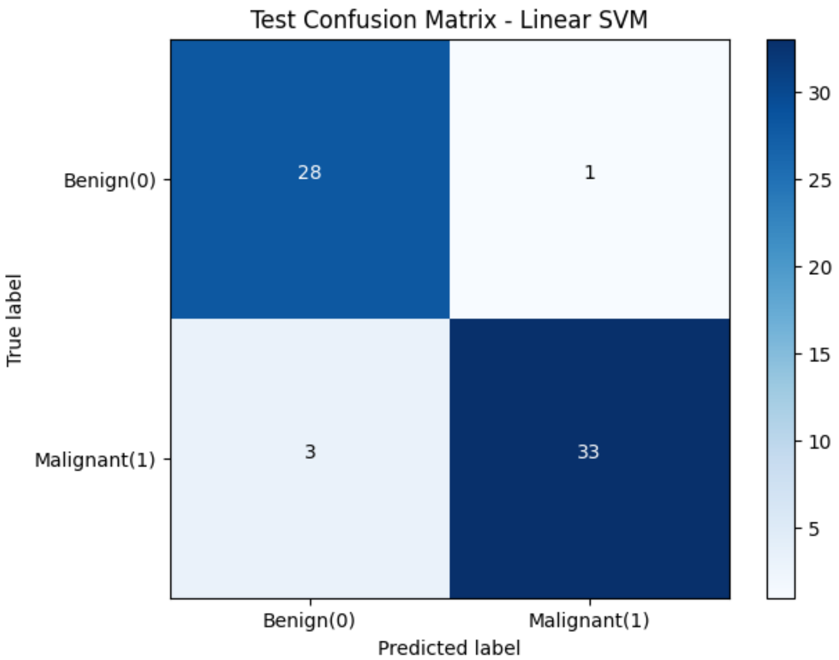
Result: SVMs for Ovarian Cancer Detection

Task – 1 : Train and Test dataset: Split data into training (70%) and testing (30%)

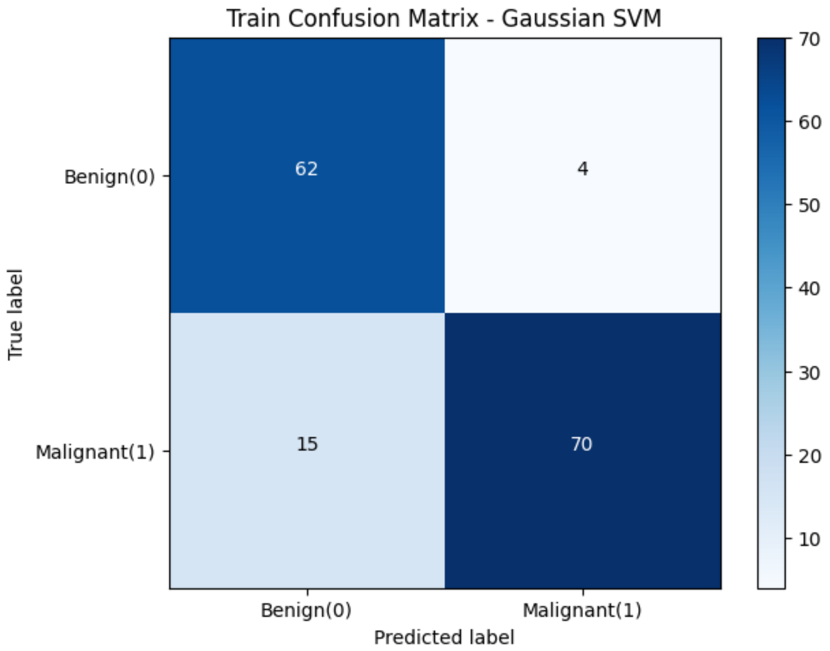
- **Linear kernel SVM:**
  - Confusion matrix for Train-Linear :



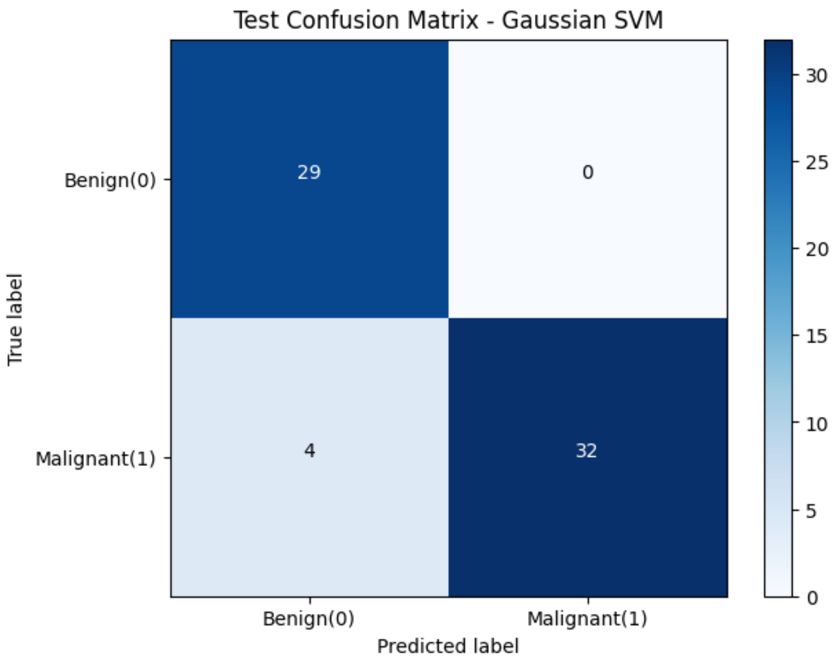
  - Confusion matrix for Test-Linear :



- **Gaussian (RBF) kernel SVM:**
  - Confusion matrix for Train-Gaussian:



- Confusion matrix for Test-Gaussian:



Accuracy : Train and Test

- **Linear SVM :**
  - Train Accuracy: 0.9470,
  - Test Accuracy: 0.9385
- **Gaussian SVM :**
  - Train Accuracy: 0.8742,
  - Test Accuracy: 0.9385

Generalization Capability

- **Linear SVM:**

- The small gap between training and testing accuracies (94.70 % − 93.85 % = 0.85 %) indicates that the model generalizes well to unseen data.
- Consistent performance across training and testing datasets demonstrates the model's robustness and ability to balance bias and variance.

- **Gaussian SVM:**
  - The larger gap between training and testing accuracies (93.85 % − 87.42 % = 6.43 %) suggests underfitting during training.
  - Although the test accuracy matches the Linear SVM, this is likely due to coincidence rather than effective learning.

Conclusion of Task-1:

- **Better Generalization**: The Linear SVM demonstrates better generalization capability due to its consistent performance across training and testing datasets.
- **Gaussian SVM:** While achieving the same test accuracy, it underfits during training, making it less reliable in practice.

Task – 2 : Change Box Constraint : Using Box Constraint values = { 0.1, 1, 10}

Results and Observations:

- **Box Constraint C = 0.1**
  - Linear Kernel Confusion Matrix (Test):

$$\begin{bmatrix} 29 & 0 \\ 4 & 32 \end{bmatrix}$$

  - Gaussian Kernel Confusion Matrix (Test):

$$\begin{bmatrix} 27 & 2 \\ 5 & 31 \end{bmatrix}$$

  Observation: At low C both models tend to prioritize the margin over exact classification, leading to more misclassifications in the Gaussian kernel compared to the linear kernel.

- **Box Constraint C = 1**
  - Linear Kernel Confusion Matrix (Test) :

$$\begin{bmatrix} 28 & 1 \\ 3 & 33 \end{bmatrix}$$

  - Gaussian Kernel Confusion Matrix (Test) :

$$\begin{bmatrix} 29 & 0 \\ 4 & 32 \end{bmatrix}$$

  Observation: Both kernels improve classification accuracy. The Gaussian kernel achieves perfect classification for one class (column 2), while the linear kernel balances misclassifications across both classes.

- **Box Constraint C = 10**
  - Linear Kernel Confusion Matrix (Test)

$$\begin{bmatrix} 29 & 0 \\ 2 & 32 \end{bmatrix}$$

  - Gaussian Kernel Confusion Matrix (Test) :

$$\begin{bmatrix} 28 & 1 \\ 4 & 32 \end{bmatrix}$$

Observation: The linear kernel achieves the highest accuracy at this C, while the Gaussian kernel shows slight overfitting with a marginal drop in performance.

Analysis of Box Constraint (C) on Linear and Gaussian SVMs

1. Linear Kernel SVM:

- As C increases, the linear kernel consistently improves performance, with the best results at C=10. This kernel generalizes well, even at higher C, balancing the margin and classification accuracy.

2. Gaussian Kernel SVM:

- The Gaussian kernel is sensitive to changes in C, It performs well at C=1, but overfitting becomes apparent at C=10, leading to decreased generalization.

Comparison:

- At C=1, the Gaussian kernel achieves near-perfect classification for the first class.
- At C=10, the linear kernel outperforms the Gaussian kernel, showcasing better generalization for this dataset.

Conclusion of Task - 2 :

- Best Kernel : The Linear kernel with C=10 exhibits the best balance between precision and recall across classes.
- Observations on C :
    - Smaller C values emphasize wider margins but can lead to underfitting.
    - Larger C values prioritize minimizing classification errors but can lead to overfitting in the Gaussian kernel.

Task – 3 : Performance of Voting Ensemble:

- Train Accuracy: 94.04%
- Test Accuracy: 93.85%

The voting ensemble successfully combines the strengths of individual classifiers, yielding robust performance across both the training and testing datasets.

Confusion Matrices Comparison:

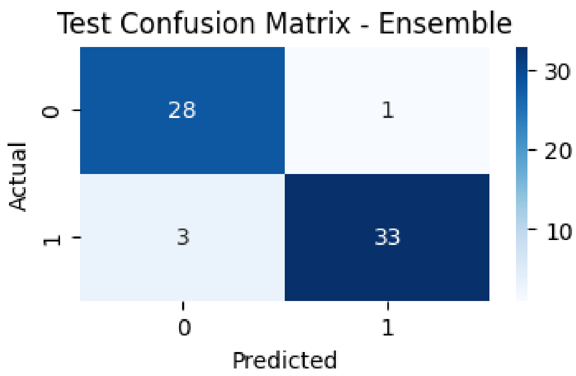
1. **Voting Ensemble:**
    - Train Confusion Matrix:

$$\begin{bmatrix} 62 & 4 \\ 5 & 80 \end{bmatrix}$$

Misclassifications are minimal, indicating the model effectively balances learning patterns without overfitting.

    - Test Confusion Matrix:

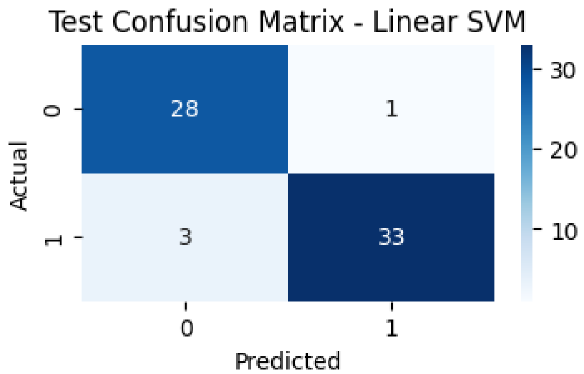
$$\begin{bmatrix} 24 & 1 \\ 3 & 33 \end{bmatrix}$$



The ensemble generalizes well, with low misclassification rates in both classes.

2. **Linear SVM:**
   - Test Confusion Matrix:

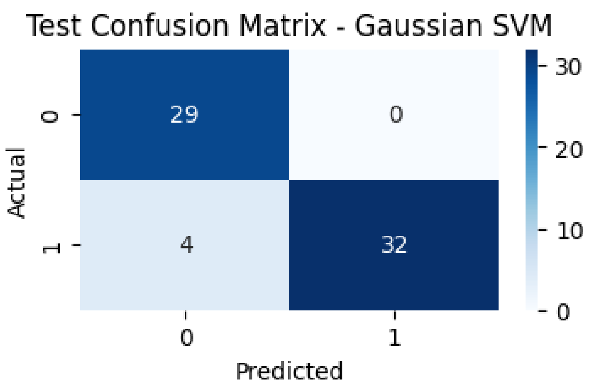
$$\begin{bmatrix} 28 & 1 \\ 3 & 33 \end{bmatrix}$$



Matches the ensemble results, demonstrating consistent classification performance.

3. **Gaussian SVM:**
   - Test Confusion Matrix:

$$\begin{bmatrix} 29 & 0 \\ 4 & 32 \end{bmatrix}$$



While the Gaussian SVM achieves perfect classification of class 0, it introduces one additional misclassification in class 1 compared to the ensemble.

Observations:

- The voting ensemble provides a balanced approach to classification, reducing misclassifications compared to the Gaussian SVM and matching the performance of the linear SVM.
- Its ability to maintain high accuracy across both datasets highlights its robustness and generalization capability.

Conclusion of Task – 3 :

The results demonstrate that ensemble learning effectively improves classification performance by combining the strengths of individual models. The voting ensemble achieves consistent and reliable outcomes, making it a preferable choice for applications requiring robust generalization.