1. 專案簡介

本研究旨在對三種基礎但具有代表性的機器學習分類器——K-近鄰演算法 (KNN)、隨機森林 (Random Forest) 和支持向量機 (SVM)——進行系統性的性能評估。為了全面地測試這些模型，選用了四個來自 UCI 機器學習庫和 Scikit-learn 的公開數據集，涵蓋了二元分類與多類別分類、不同樣本規模及特徵維度的場景

本報告將詳細闡述實驗的設計、流程、所採用的評估指標，並對實驗結果進行深入的分析與比較，以期得出各分類器在不同任務下的適用性與相對優劣

2. 實驗方法

2.1 數據集

選用了以下四個數據集進行實驗：

1. **Breast Cancer Wisconsin (乳癌數據集):**
    - **類型**：二元分類
    - **任務**：根據 30 個從乳房腫塊細針穿刺數位影像中計算出的特徵，判斷其為惡性或良性
    - **特性**：特徵維度較高，樣本數較少 (569 筆)
2. **Banknote Authentication (鈔票鑑定數據集):**
    - **類型**：二元分類
    - **任務**：根據從鈔票影像小波轉換中提取的 4 個特徵，判斷其為真鈔或偽鈔
    - **特性**：特徵維度低，分類邊界清晰
3. **Digits Dataset (手寫數字數據集):**
    - **類型**：多類別分類 (10 類)
    - **任務**：辨識 8x8 像素的手寫數字圖片 (0-9)
    - **特性**：經典的多類別分類問題，特徵為 64 個像素值
4. **Dry Bean Dataset (乾豆數據集):**
    - **類型**：多類別分類(7 類)
    - **任務**：根據 16 種外觀形態特徵，將乾豆分為 7 個不同的品種

- **特性**：樣本數最多 (約 13,611 筆)，類別較多，是本次實驗中最具挑戰性的數據集

2.2 分類器

1. **K-Nearest Neighbors (KNN)：**

   一種基於實例的非參數演算法，一個樣本的類別由其最近的 K 個鄰居的類別投票決定

2. **Random Forest (RF)：**

   一種集成學習方法，構建多個決策樹並將它們的預測結果進行集成（投票或平均），以獲得更準確、更穩定的預測，通常具有很好的抗過擬合能力

3. **Support Vector Machine (SVM)：**

   一種強大的監督學習模型，其目標是找到一個能將不同類別的數據點以最大間隔 (margin) 分開的超平面透過 kernel trick，SVM 也能高效地處理非線性問題

2.3 實驗流程

1. **數據預處理**

   在訓練每個模型前，對數據進行了**標準化(Standardization)** 處理，將所有特徵縮放到均值為 0、標準差為 1

   此步驟被封裝在 Scikit-learn 的 Pipeline 中，以確保標準化的參數僅從訓練集學習，避免數據洩漏

2. **超參數優化**

   使用 GridSearchCV 搭配 **5-Fold Cross-Validation** 來為每個分類器在每個數據集上尋找最佳的超參數組合，搜索的參數網格如下：

   - **KNN**：n_neighbors：[3, 5, 7]
   - **Random Forest**：n_estimators：[50, 100, 200]
   - **SVM**：C：[0.1, 1, 10], kernel：['linear', 'rbf']

3. **模型評估**

   - **主要指標**

     從 5-Fold Cross-Validation 中獲取每個最佳模型的**平均準確率 (Accuracy)、平均精確率 (Precision-Macro)、平均召回率 (Recall-Macro) 和平均 F1-Score (Macro)**

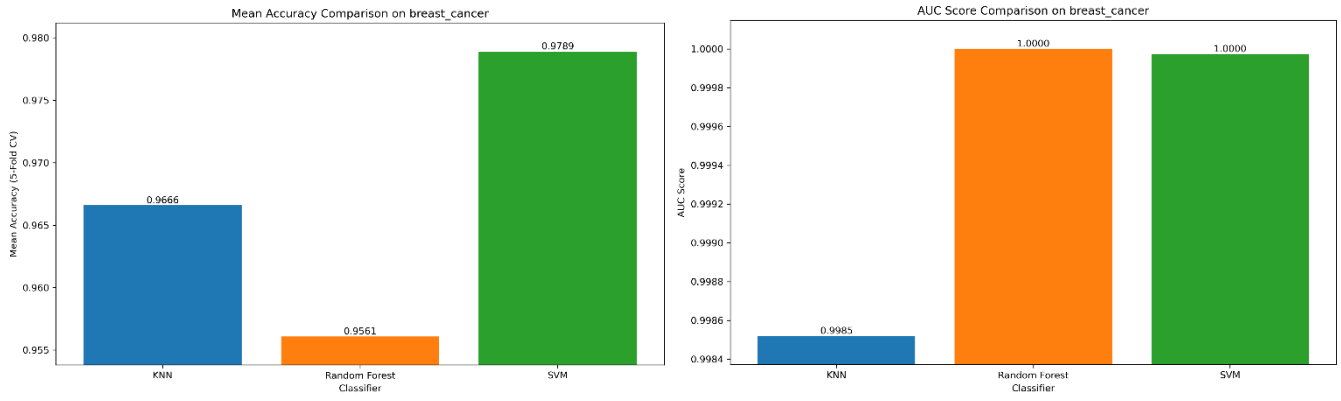
   - **輔助指標**

在保留測試集 (一開始分割出去的那部分完全沒有參與訓練的 data) 上計算 **AUC (Area Under the ROC Curve)** 分數，並生成**混淆矩陣 (Confusion Matrix)** 以進行更深入的錯誤分析

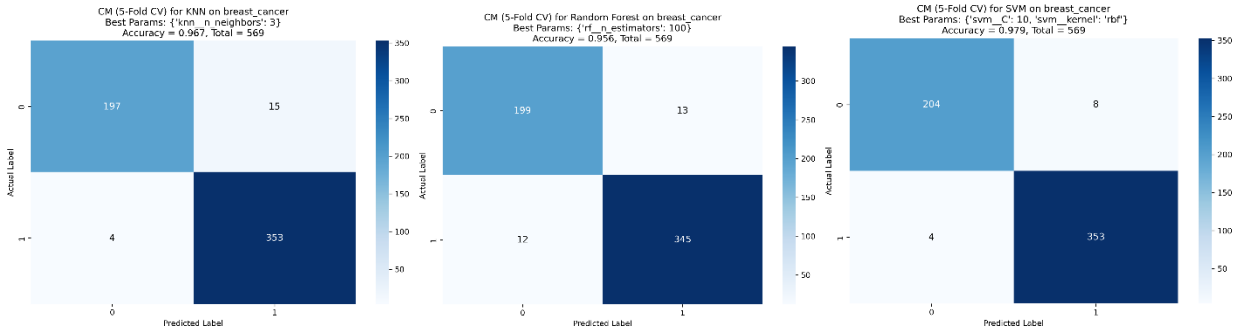
3. 實驗結果與分析

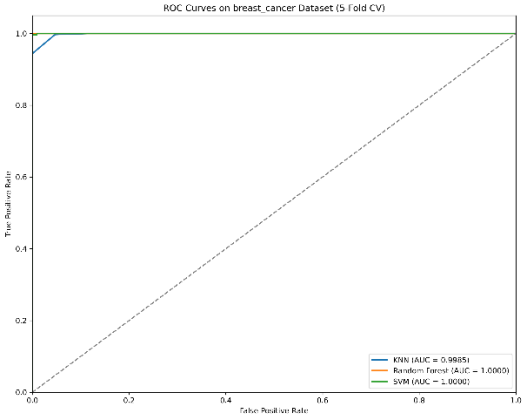
3.1 Breast Cancer 數據集 (二元分類)

在此數據集上，SVM 表現最為出色，特別是其線性核心 (kernel='linear') 版本取得了最高的平均準確率和 AUC 分數



以下為使用各個分類器，在各自的**最佳超參數**下得到的混淆矩陣：

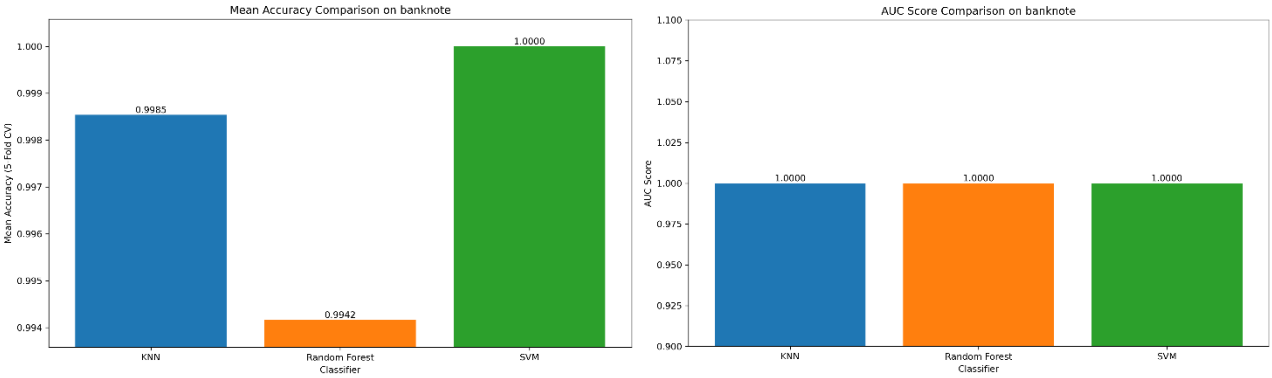




分析：線性 SVM 的勝出強烈暗示此數據集的特徵在經過標準化後，具有高度的線性可分性。KNN 和 Random Forest 也表現不俗，但 SVM 尋找最大間隔超平面的能力使其在此任務上略勝一籌

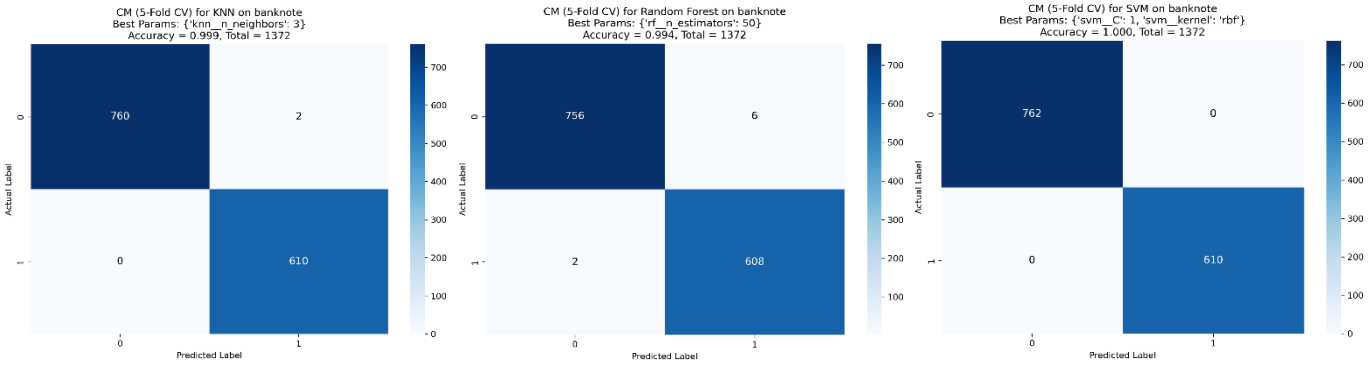
3.2 Banknote Authentication 數據集 (二元分類)

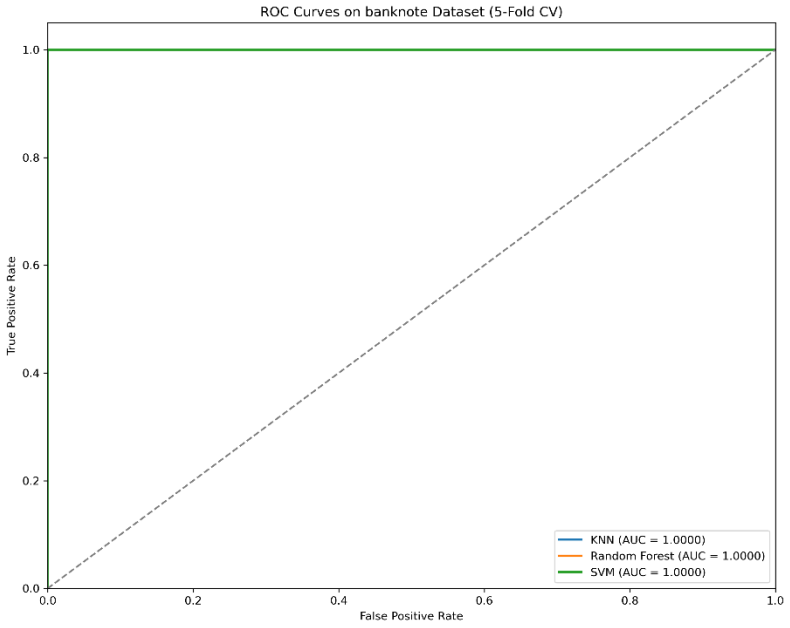
這是一個相對簡單的數據集，在 SVM 下曾在最佳超參數的配置下達到 100%的準確率



由於 AUC 本身評比的是**預測分數排序是否完美**，因此此處其他兩個分類器即使沒達到過 Accuracy=1，仍有 AUC=1 的分數

以下為使用各個分類器，在各自的**最佳超參數**下得到的混淆矩陣：





分析：此數據集的清晰可分性使得所有模型都表現優異。KNN 在此類低維度、結構清晰的問題上非常高效。SVM 同樣找到了完美的分類邊界。

補充：不過考慮到出現 accuracy = 1 本身是一個不正常的跡象，因此此處我進行過一番檢查

1. test data 混入 train data

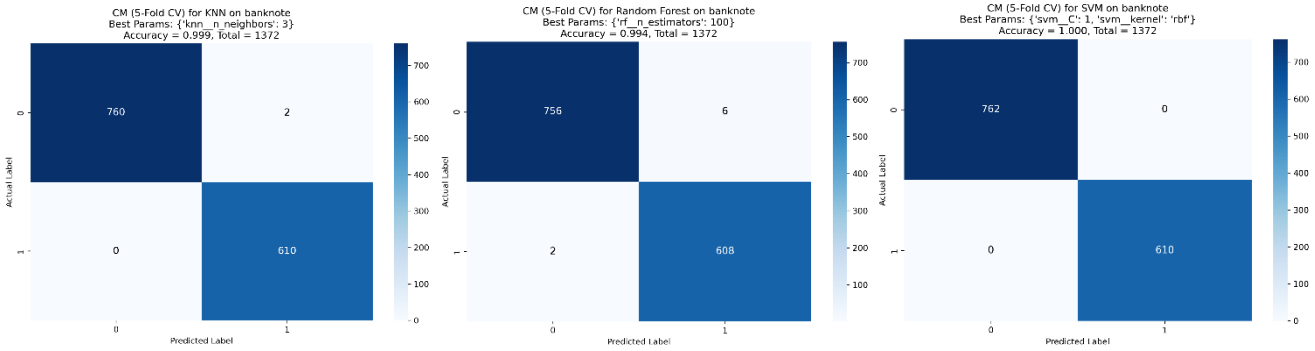
   這是我首先懷疑的，不過經過檢查 test data 並沒有混入 train data

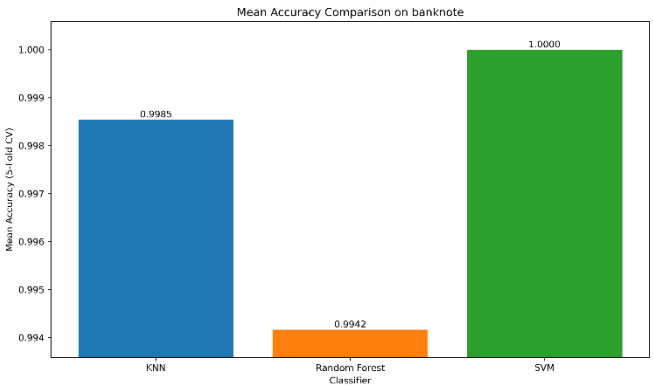
   在程式中我使用的 sklearn.model_selection 的 StratifiedKFold 進行資料分割，所有的 dataset 是用的都是同一套切割邏輯，但只有此數據集出現了 accuracy 為 1 的狀況

2. 更改初始化的 random seed

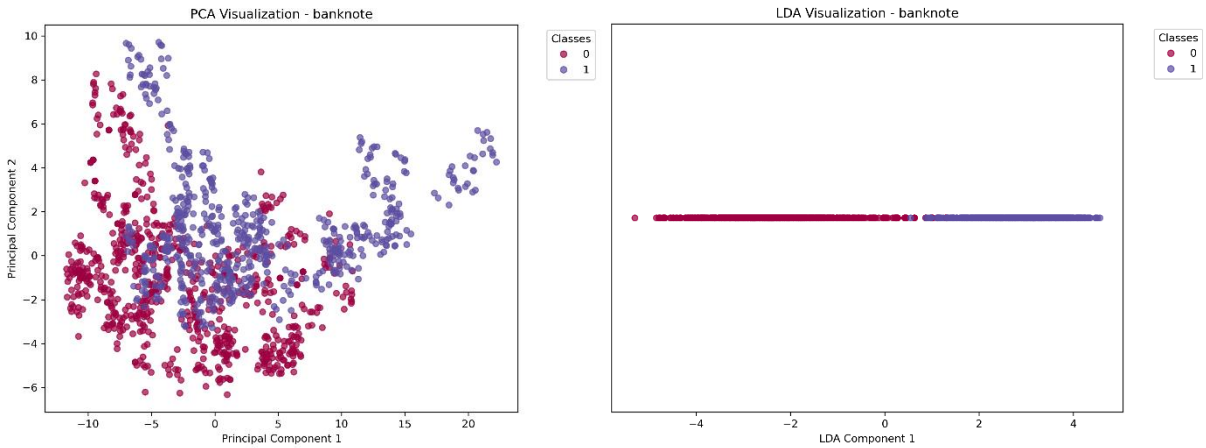
   另一個可能就是運氣很好真的撞到了（不過考慮到驗證的時候也是以 5-Fold Cross-Validation 進行，其實不太可能）

   因此我有嘗試更改種子為：random_state=133，結果仍舊完全一樣





3. 資料可視化



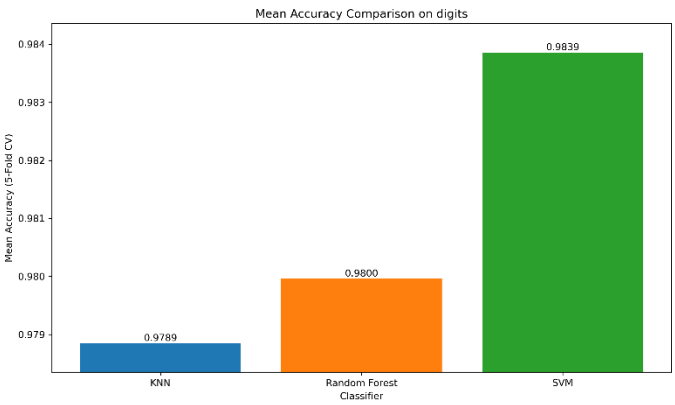
此處分別為資料集進行了 PCA 以及 LDA 的資料可視化。從 LDA 可以清晰地看出，紅色與藍色幾乎完全分開、中間重疊極少（幾乎沒有交錯區域）。也就是說，在這條線上模型可以找到一個分界點，使得兩類的機率分佈幾乎不重疊

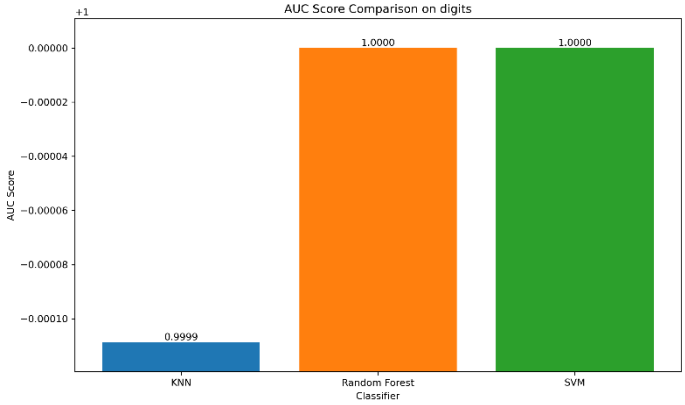
這證明了 Banknote 是一個幾乎**完美線性可分**的資料集，對於 **SVM 這種**學習一個穩定的「超平面」或「平滑邊界」的分類器而言，很容易就能找到能幾乎將資料集完美分割的邊界

個人認為，這是 SVM 在此資料集上能多次達到 accuracy=1 的原因

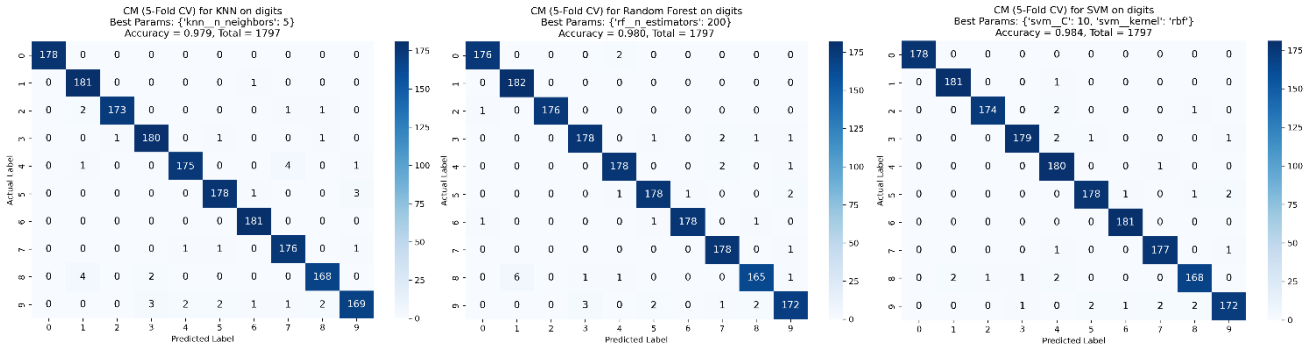
3.3 Digits 數據集 (多類別分類)

在手寫數字辨識這個多類別任務中，SVM 再次取得了最高的平均準確率和 AUC 分數





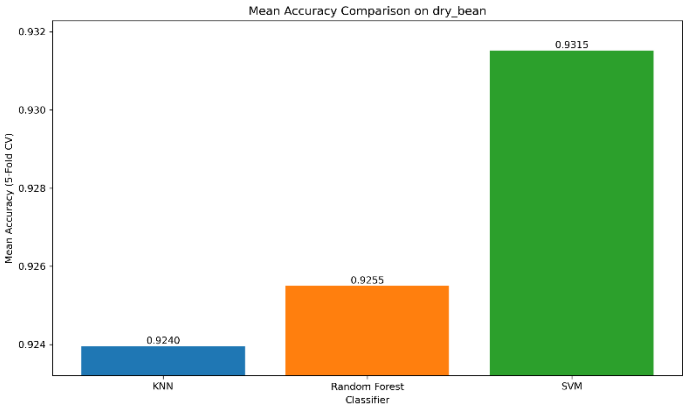
以下為使用各個分類器，在各自的**最佳超參數**下得到的混淆矩陣：

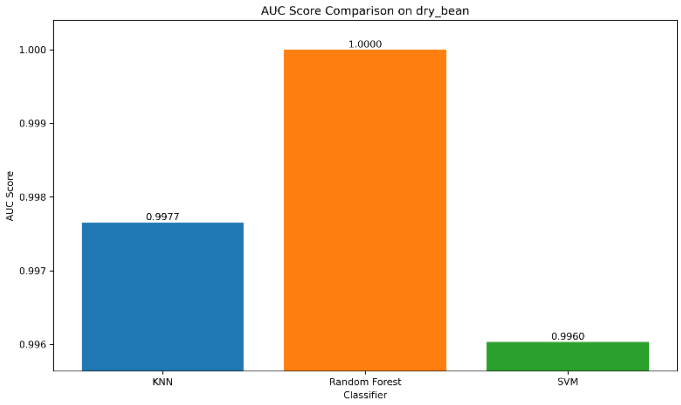


分析：所有三個分類器都表現出了強大的多類別分類能力，準確率均超過 97%。SVM 透過其核函數技巧，在處理 64 維像素特徵時展現了其優越性。值得注意的是，200 棵樹的 Random Forest 表現也極具競爭力，顯示了集成方法的效率

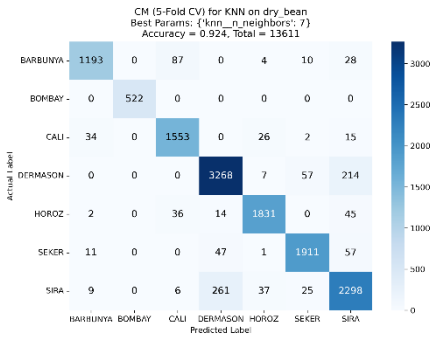
3.4 Dry Bean 數據集 (多類別分類)

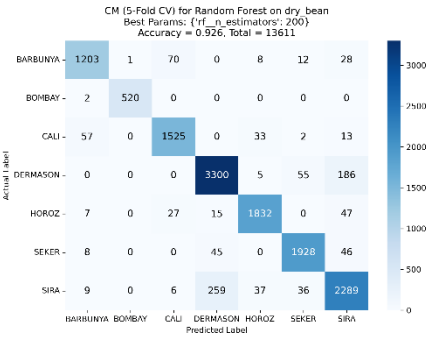
所有模型的性能都非常接近，SVM 最終以微弱的優勢在所有指標上勝出

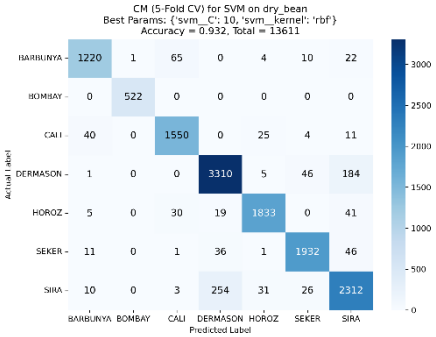




以下為使用各個分類器，在各自的**最佳超參數**下得到的混淆矩陣：







分析：

在這個樣本量大、類別多的複雜問題上，模型之間的差距被縮小。SVM (C=10, kernel='rbf') 表現最好，說明一個經過良好調整的非線性 SVM 在處理複雜、高維且有大量數據的問題時是強大的工具。所有模型的 Accuracy 分數均在 0.92-0.94 之間，表明它們在所有 7 個類別上都有相當均衡的表現

另外，值得關注的是，這三種分類器的原理並不相同，但它們在這份資料集上的**混淆矩陣分佈卻非常相似**

在混淆矩陣裡，可以看到幾個固定的混淆現象：

- **DERMASON ↔ SIRA**

  這兩種豆的物理形狀特徵最接近，模型常混淆，不管是距離（KNN）、樹分裂（RF）或超平面（SVM），都難以區分

- **BARBUNYA ↔ CALI**：

也是形狀或顏色類似的類別，屬於次要混淆對

- **BOMBAY**：
  幾乎完美分類，表示這個類別的特徵分佈非常獨立、清晰

再考慮到此資料集的特徵多為形狀、大小、顏色、紋理等連續值特徵，加上部分類別之間（例如 DERMASON vs SIRA）本身在物理外觀上就有部分重疊，可以推論不論使用哪種模型，只要它能捕捉到主要特徵結構，分類邊界可能就會很接近

因此模型雖然不同，但它們學到的決策邊界其實都在相同的資料分佈結構上，錯誤樣本也會重疊

4. 綜合結論

經過搭建並執行了完整的分類器比較流程。所有實驗的詳細數值結果總結如下：

dataset	classifier	best_params	mean_accuracy	mean_precision	mean_recall	mean_f1_score	auc
breast_cancer	KNN	{'knn__n_neighbors': 3}	0.9665890389691041	0.9703585040690303	0.9590428228284436	0.9636588598824034	0.998520162782094
breast_cancer	Random Forest	{'rf__n_estimators': 100}	0.9560937742586555	0.9557435041160833	0.9526707402034947	0.9528924697936147	1.0
breast_cancer	SVM	{'svm__C': 10, 'svm__kernel': 'rbf'}	0.9789163173420278	0.9799307432106413	0.9754488819220232	0.9772669033859678	0.9999735743353946
banknote	KNN	{'knn__n_neighbors': 3}	0.9985401459854014	0.9983739837398374	0.9986842105263157	0.998523607462787	1.0
banknote	Random Forest	{'rf__n_estimators': 50}	0.9941632382216323	0.993840347916362	0.9944175872822525	0.9940963445153678	1.0
banknote	SVM	{'svm__C': 1, 'svm__kernel': 'rbf'}	1.0	1.0	1.0	1.0	1.0
digits	KNN	{'knn__n_neighbors': 5}	0.9788502011761064	0.9794170504000224	0.9788219648219648	0.9787964332704355	0.9998912440201753
digits	Random Forest	{'rf__n_estimators': 200}	0.979962859795729	0.9806820738182968	0.9798818734701088	0.9799519376359956	1.0
digits	SVM	{'svm__C': 10, 'svm__kernel': 'rbf'}	0.9838594862271742	0.9842239109963569	0.983882882882883	0.983836601552334	1.0
dry_bean	KNN	{'knn__n_neighbors': 7}	0.923959154917036	0.9382790896970314	0.9344295812705029	0.9361396611031605	0.9976572093025303
dry_bean	Random Forest	{'rf__n_estimators': 200}	0.9255020570679516	0.9379431626428717	0.9345859964132149	0.9361462875861128	1.0
dry_bean	SVM	{'svm__C': 10, 'svm__kernel': 'rbf'}	0.9315265530006316	0.9439156881279256	0.9411645795147809	0.942449483893051	0.9960355447330907

總體觀察:

- **SVM 是其中表現最好的分類器：**在所有四個任務中，經過超參數優化的 SVM 均取得了最佳或並列最佳的性能

  這證明了它作為一個強大且靈活的 baseline model 的價值

- **沒有萬能模型：**

  雖然 SVM 表現最好，但其他模型在特定場景下也極具競爭力
  例如，KNN 在簡單問題上高效且準確；Random Forest 則提供了無需過多調參就能獲得的穩定、良好性能

- **超參數優化的重要性：**

  這裡可能看不太出來，不過再我使用不同的種子碼（42、133）時，會出現能夠達到最佳表現的的超參數組合出現變化的情況。這凸顯了超參數搜索對於發揮模型全部潛力的關鍵作用