PAPER REPORT - MACHINE LEARNING PROJECT 1
Student: Imbrea Giulia Stefania-343C4

Introduction

In the field of machine learning, practitioners and researchers frequently encounter three essential tasks:

1. **Data Visualization and Exploration (Exploratory Data Analysis)**: Understanding the underlying structure and characteristics of the data.

2. **Feature Extraction**: Identifying relevant attributes or features from the data to be used in the analysis, such as classification, regression, or anomaly detection tasks.

3. **Model Evaluation**: Assessing multiple models to determine the best solution for a given problem.

This project focuses on applying these tasks to image classification problems using two popular datasets: **Fashion-MNIST** and **Fruits-360**. The goal is to apply feature extraction techniques to these image datasets and evaluate the performance of various machine learning classifiers.

The **Fashion-MNIST** dataset consists of 70,000 grayscale images, each 32x32 pixels, representing 10 different types of clothing items, such as t-shirts, pants, and boots. The **Fruits-360** dataset contains around 55,000 RGB images, representing 80 different types of individual fruits. The objective for both datasets is to classify the images into the correct categories using machine learning algorithms.

The primary tasks of this report are:

- **Feature Extraction**: To identify and extract meaningful features from images using two methods: **Histogram of Oriented Gradients (HOG)** and **Principal Component Analysis (PCA)**.

- **Model Classification**: To train and evaluate several machine learning algorithms, including Logistic Regression, Support Vector Machines (SVM), Random Forest, and Gradient Boosted Trees, to perform the classification task.

- **Model Optimization**: Hyperparameters of the models will be optimized using techniques like Grid Search and Cross-Validation to improve classification performance.
  This report will detail the methods and results of each of these steps, providing insights into the utility of the feature extraction methods and their impact on the classification results.

Dataset Overview

Two image datasets were used for this project: **Fashion-MNIST** and **Fruits-360**. Both datasets are publicly available and commonly used for image classification tasks in machine learning.

- **Fashion-MNIST**

  Fashion-MNIST is a dataset of 70,000 grayscale images, each sized 28x28 pixels, categorized into 10 classes representing different clothing items. The classes include:
  - T-shirt/top
  - Trouser
  - Pullover
  - Dress
  - Coat
  - Sandal
  - Shirt
  - Sneaker
  - Bag
  - Ankle boot

  This dataset has already been split into training and testing sets. The task is to classify each image into one of the 10 predefined categories.

- **Fruits-360**

  Fruits-360 is a dataset containing approximately 55,000 images, each representing a single fruit. The images are in RGB color format, and the dataset includes 80 fruit classes. Examples of classes include:
  - Apple
  - Banana
  - Orange
  - Pineapple
  - Grapes
  - Strawberry

  Similarly to Fashion-MNIST, this dataset is divided into training and testing sets. The goal is to correctly classify each fruit image into one of the 80 classes.
  Both datasets are well-suited for machine learning tasks and provide a challenge for evaluating feature extraction and classification algorithms.

Class Balance Analysis

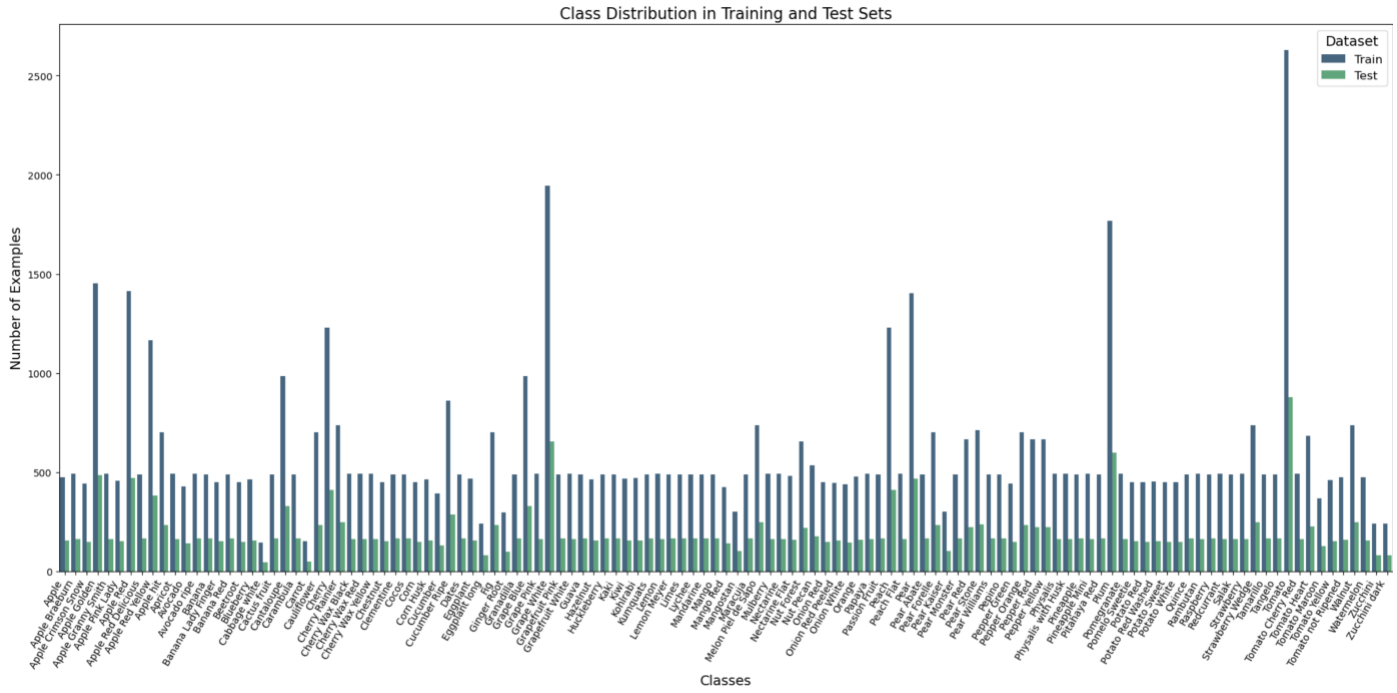
- **Fashion-MNIST Class Balance**

For the **Fashion-MNIST** dataset, we observe that the class distribution in both the training and testing sets is highly balanced. The dataset contains **10 classes** representing different types of clothing, and each class has a similar number of training and testing samples. Specifically:

- **Training set**: 6000 samples per class.
- **Testing set**: 1000 samples per class.

Each class has a similar number of examples, making the dataset ideal for training machine learning models that can generalize well across all clothing categories.

- **Fruits Dataset ClassBalance**

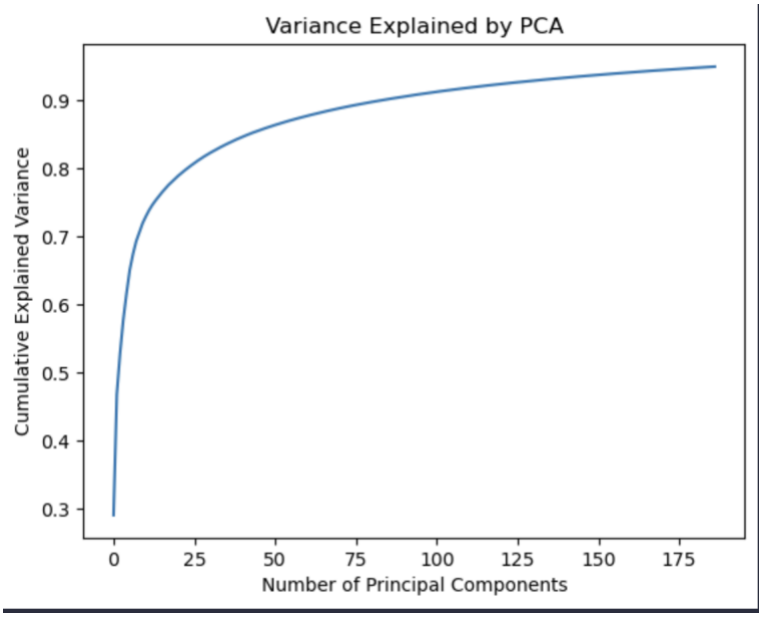


For the **Fruits-360** dataset, the class distribution is less balanced compared to Fashion-MNIST. This dataset contains **121 fruit classes**, and the number of samples per class varies significantly. Some

classes, like **Tomato**, **Grape White**, and **Plum**, have a much higher number of samples, while other classes are underrepresented.

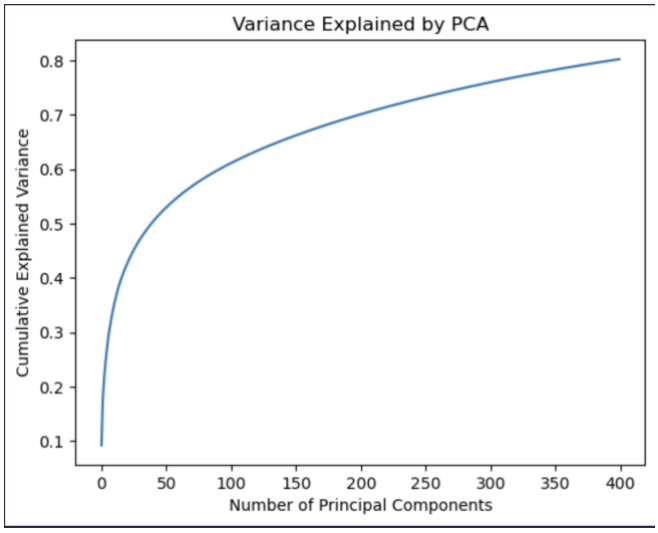
Quantitative PCA Visualization: The Degree of Cumulative Variance Explained by the Selected Number of Principal Components

- **Fashion-MNIST: Number of Selected Principal Components = 187 (95% variance retention)**



- **Explanation**: For **Fashion-MNIST**, PCA was used to reduce the dimensionality of the dataset while retaining **95% of the variance**. As a result, **187 principal components** were chosen, which captured the most important features of the images, such as edges, textures, and basic shapes.
- **Interpretation**: The plot showing the **cumulative variance explained by the selected components** will illustrate how much of the original image information is retained by the first few principal components. After 187 components, the plot reaches 95%, meaning that further components do not add significant new information. The sharp increase in cumulative variance early on indicates that most of the key features in the images are captured in just a few components, while the remaining components contribute less to the overall variance.

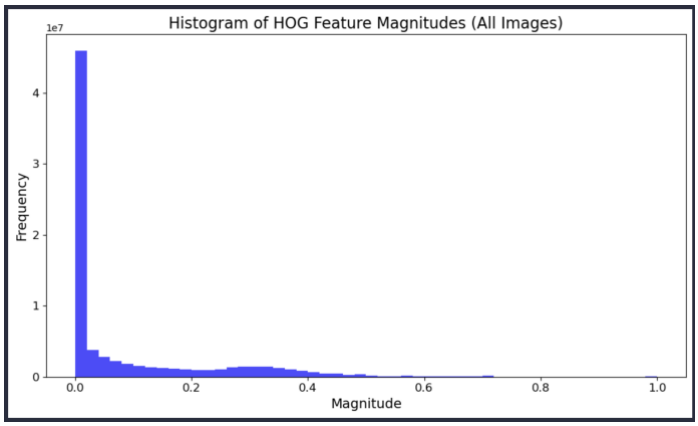
- **Fruits-360: Number of Selected Principal Components = 400 (80% variance retention)**



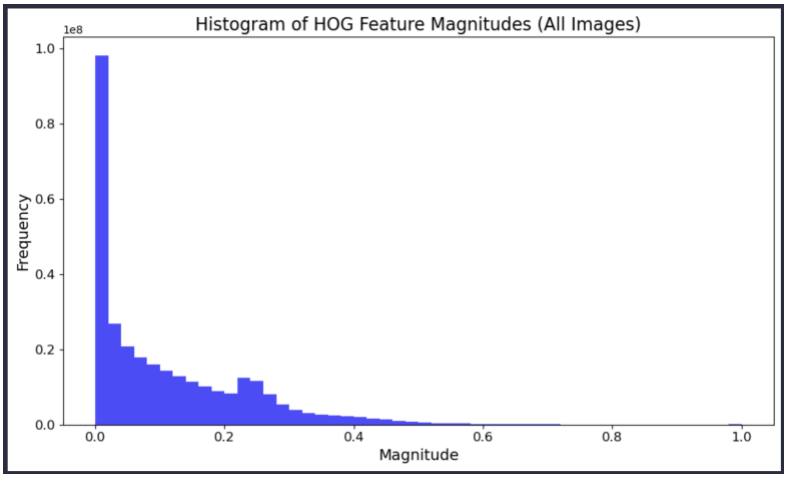
- **Explanation**: For the **Fruits-360** dataset, **PCA** was applied after extracting the **HOG features**. The goal was to reduce the dimensionality while retaining a reasonable amount of variance in the data. After applying PCA, **400 principal components** were selected, which retained **approximately 80% of the variance**. This number of components was chosen because it offered a good balance between reducing dimensionality and retaining enough information to preserve the distinguishing characteristics of the fruits.
- **Interpretation**: The plot for **Fruits-360** will show that **80% of the variance** is captured with the selected 400 components. This is a lower percentage of variance retention compared to **Fashion-MNIST** (95%), which is expected because **Fruits-360** has more complex images with greater diversity in shapes, textures, and colors. The cumulative variance curve will likely flatten more gradually than for Fashion-MNIST, indicating that more components are needed to capture the key features.

Quantitative HOG Visualization: **Histogram of HOG Feature Magnitudes**

- **Fashion-MNIST**



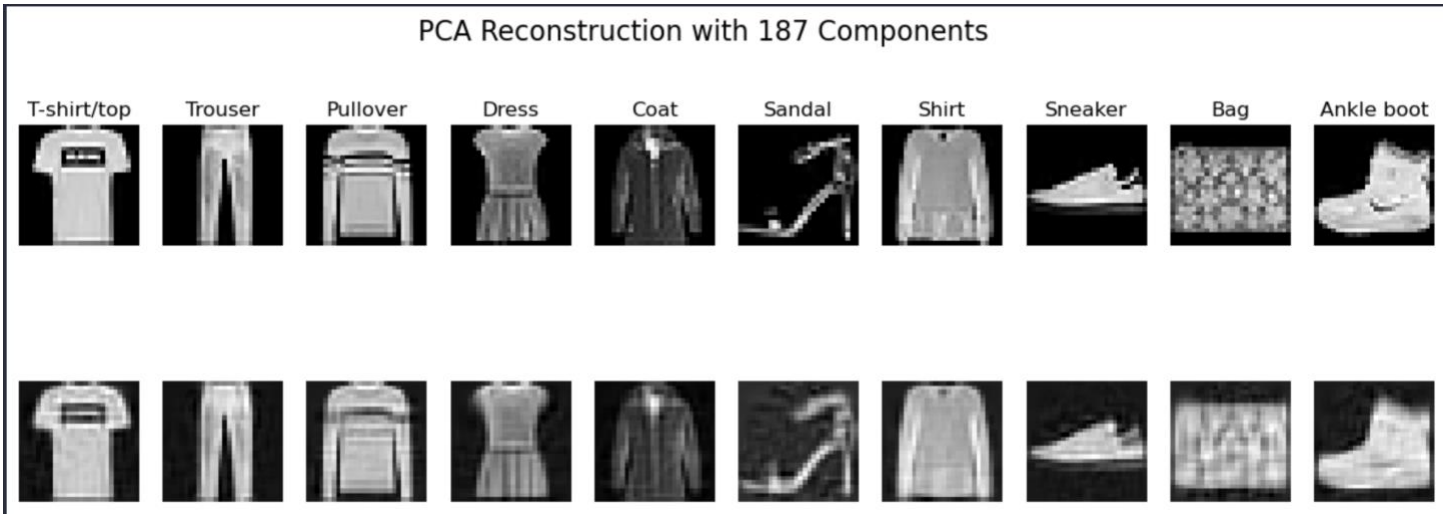
- **Fruits-360**:



The peak at lower magnitudes in both datasets suggests that a large portion of the images have **relatively uniform textures or simple backgrounds**. For Fashion-MNIST, this could correspond to areas of clothing that have smooth or less distinct edges. For Fruits-360, it may indicate the presence of simpler background textures or smoother areas of the fruit images with fewer prominent boundaries.
Qualitative visualizations

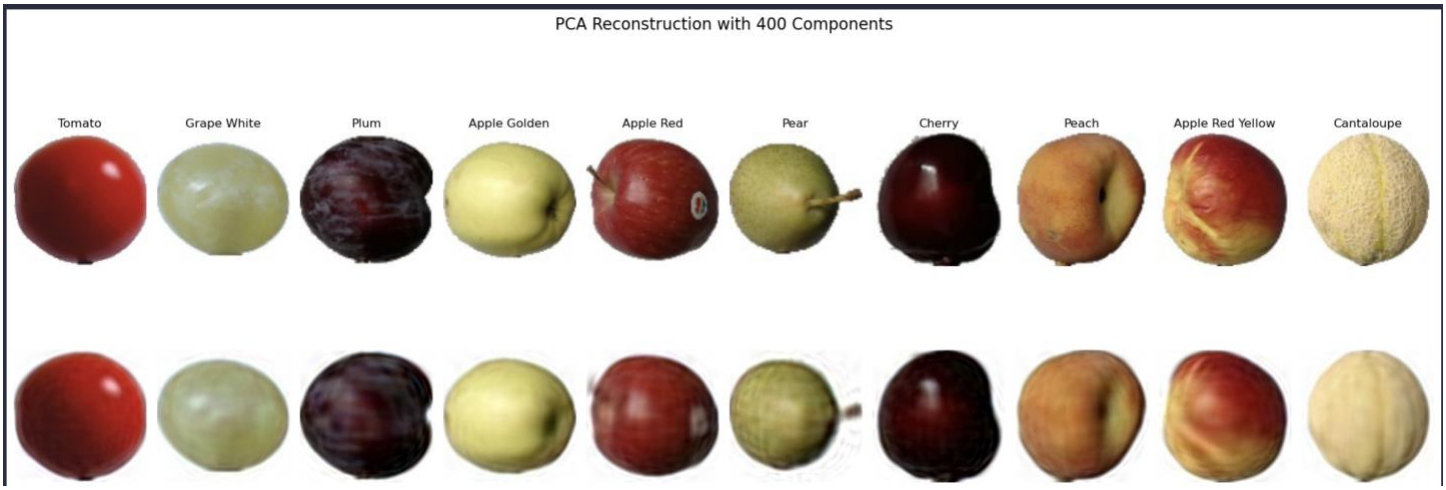
QUALITATIVE PCA VISUALIZATION – USING PCA RECONSTRUCTION

- **Fashion-MNIST**



The PCA transformation retains the most dominant features, such as the overall shape of the clothing, but reduces fine details like wrinkles and fabric texture, leading to a blurred appearance.

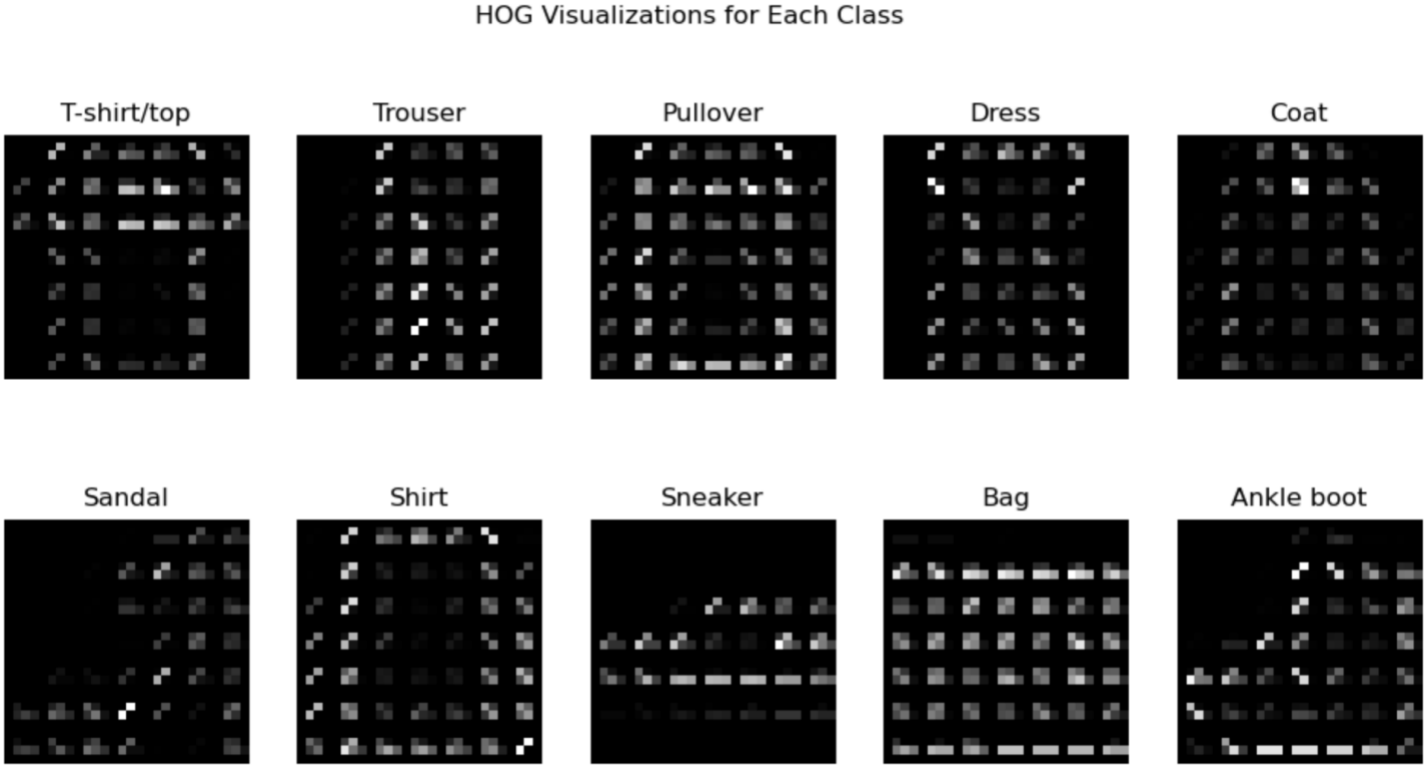
- **Fruits-360**:



Similar to the Fashion-MNIST dataset, PCA reduces the image to its most important components, removing finer details like texture or color transitions. The blurred appearance reflects the loss of complex surface details in favor of a simplified shape.

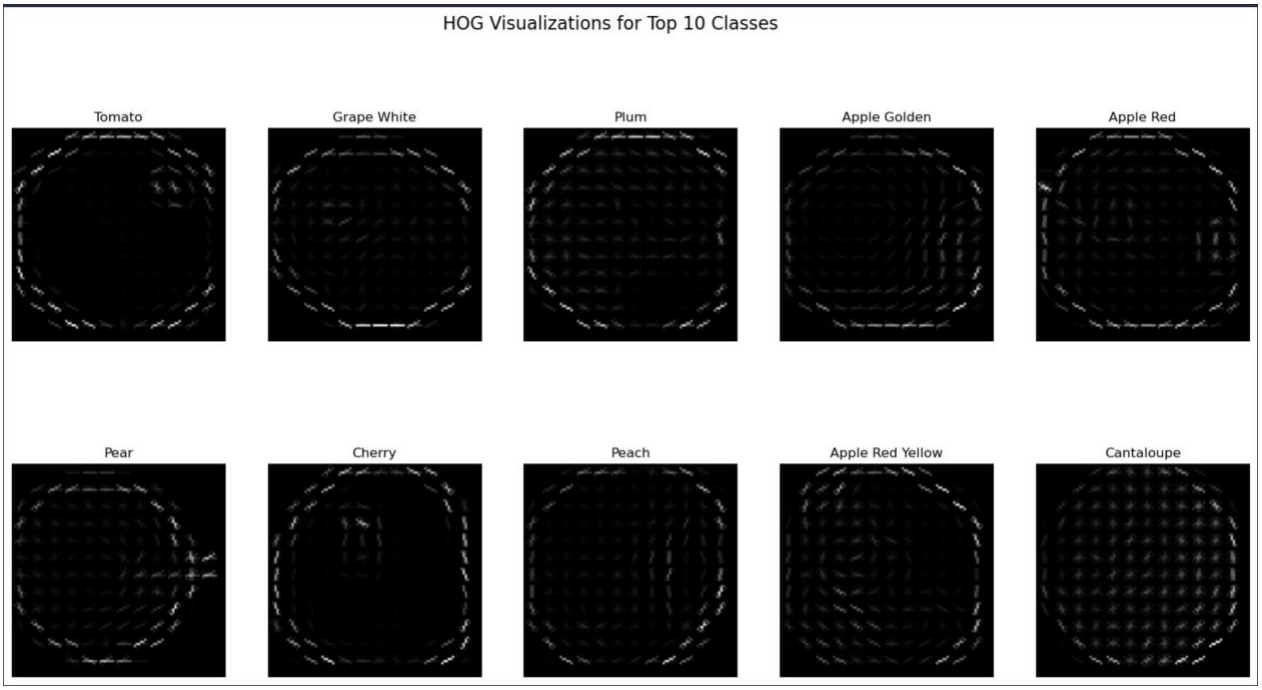
QUALITATIVE HOG VISUALIZATION

- **Fashion-MNIST**



HOG transforms the images into contour-based representations, where the shapes and boundaries of clothing items are emphasized. This transformation is very effective for capturing structural features, such as the outlines of t-shirts, pants, or boots, making it easier for the classifier to focus on shapes rather than textures.

- **Fruits-360**:



HOG accentuates the edges and contours of fruit shapes, which are important for distinguishing between different types of fruit. The high resolution of the fruit images (100x100 pixels) makes the edge detection even more effective, as it highlights the contours and boundaries of the fruits in a way that makes them easy to identify.

Feature Selection Results

Fashion-MNIST Dataset:

- Total features considered: 1483 (combined HOG and PCA).
- Features used for training: 741 features selected after applying SelectPercentile (top 50%).
- Explanation: The feature selection process removed less informative features, keeping the most relevant ones for classification based on their statistical significance.

Fruits-360 Dataset:

- Total features considered: 400 (the hog results were given as PCA input).
- Features used for training: 200 features selected after applying SelectPercentile (top 50%).
- Explanation: Similar to Fashion-MNIST, less relevant features were discarded, retaining only the features with the highest correlation to the target variable.

SelectPercentile was chosen over **Variance Threshold** because it selects features based on their **statistical relevance** to the target variable, rather than just removing features with low variance. This ensures that even features with low variance but high relevance to the classification task are preserved.

1. Logistic Regression Evaluation

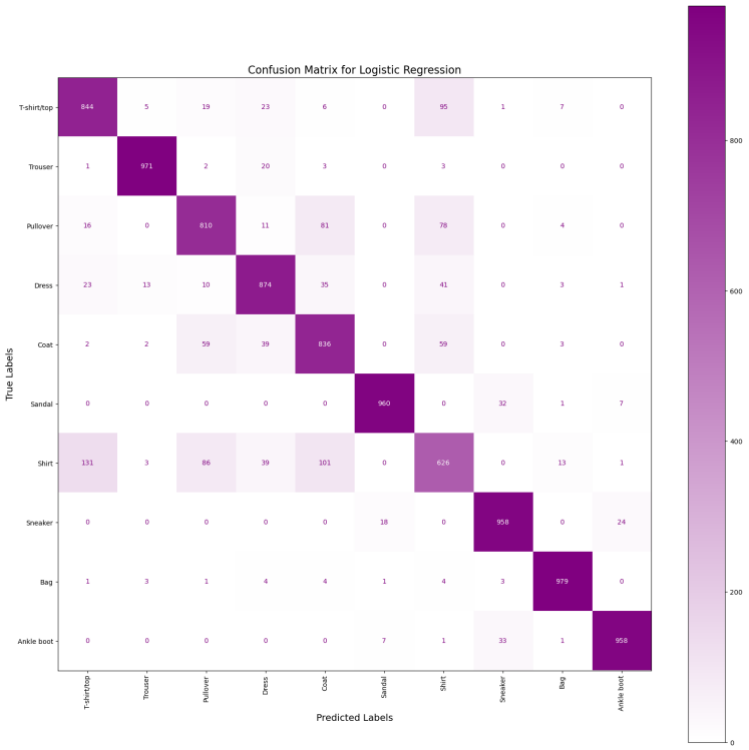
Fashion-MNIST Dataset

- Best Hyperparameters:
  - C = 0.1, multi_class = 'ovr', penalty = 'l2', solver = 'lbfgs'
- Accuracy:
  - Validation Set: 88.74%
  - Test Set: 88.16%
- Classification Report (Test Set):

Test Classification Report for Logistic Regression:

	precision	recall	f1-score	support
T-shirt/top	0.83	0.84	0.84	1000
Trouser	0.97	0.97	0.97	1000
Pullover	0.82	0.81	0.82	1000
Dress	0.87	0.87	0.87	1000
Coat	0.78	0.84	0.81	1000
Sandal	0.97	0.96	0.97	1000
Shirt	0.69	0.63	0.66	1000
Sneaker	0.93	0.96	0.95	1000
Bag	0.97	0.98	0.97	1000
Ankle boot	0.97	0.96	0.96	1000
accuracy			0.88	10000
macro avg	0.88	0.88	0.88	10000
weighted avg	0.88	0.88	0.88	10000

- Confusion Matrix



Explanation:

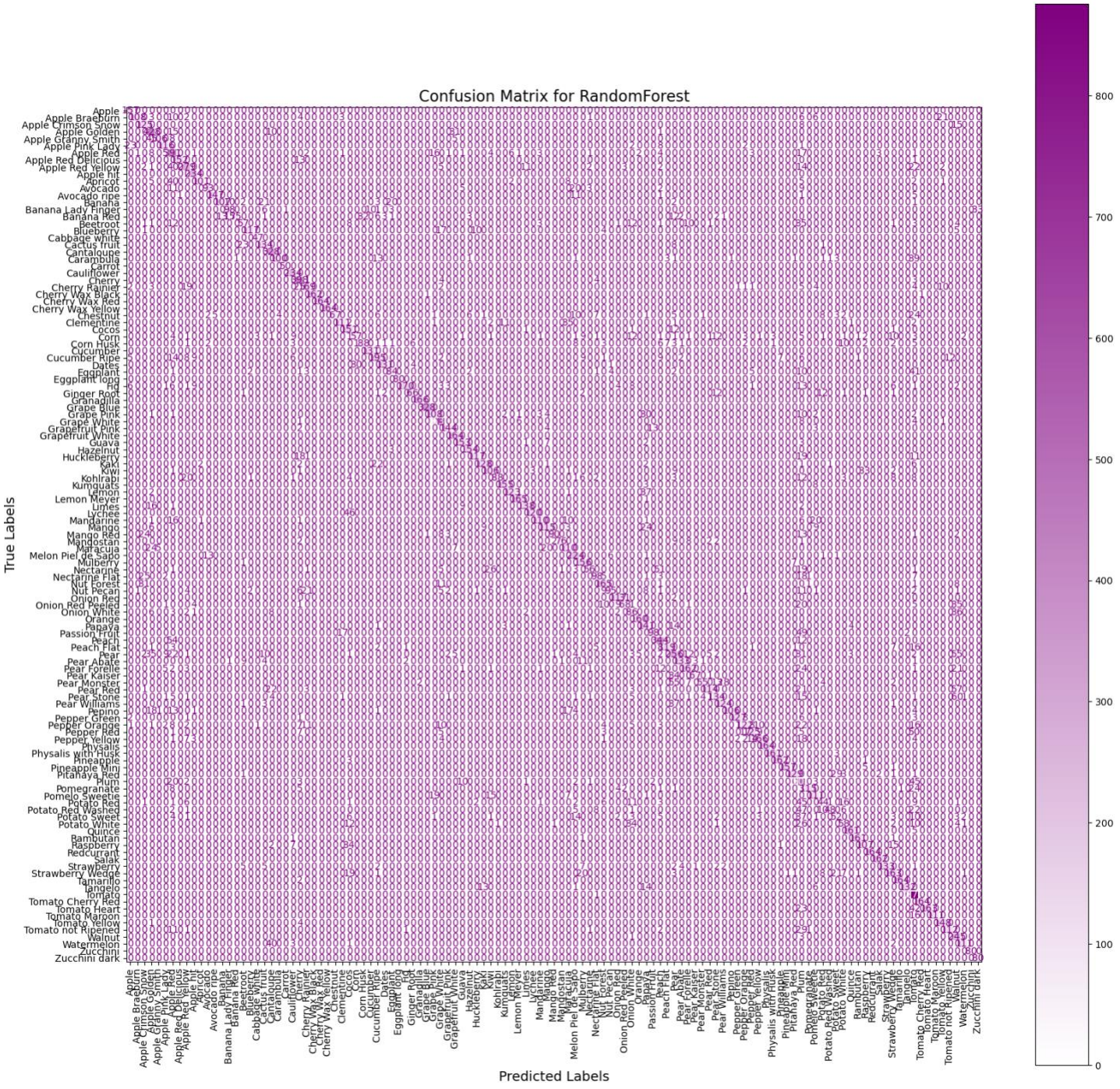
- The **best hyperparameters** were selected through a **3-fold cross-validation** process. The **accuracy** of the model on both the validation set and the test set was approximately **88%**, showing strong performance.
- **Precision, recall, and F1-score** for each class are fairly balanced, with **Trouser** and **Sandal** showing the highest values (around 0.97), while **Shirt** has a relatively lower precision and recall (0.69 and 0.63, respectively).
- The **macro avg** and **weighted avg** indicate that the model performs consistently well across all classes, but **class imbalance** might affect the performance for certain classes like **Shirt**.

Fruits-360 Dataset

- **Best Hyperparameters:**
  - **C = 1, multi_class = 'multinomial', penalty = 'l2', solver = 'lbfgs'**
- **Accuracy:**
  - **Validation Set: 99.50%**
  - **Test Set: 80.57%**
- **Classification Report (Test Set):**

```
Test Classification Report for Logistic Regression:
                     precision    recall  f1-score   support

              Apple       0.98      1.00      0.99       157
     Apple Braeburn       0.79      0.63      0.70       164
 Apple Crimson Snow       0.70      0.86      0.77       148
       Apple Golden       0.86      0.80      0.83       485
 Apple Granny Smith       0.82      0.70      0.75       164
    Apple Pink Lady       0.79      0.99      0.88       152
          Apple Red       0.61      0.68      0.64       472
Apple Red Delicious       0.95      0.96      0.95       166
   Apple Red Yellow       0.59      0.63      0.61       383
          Apple hit       0.96      1.00      0.98       234
            Apricot       0.93      0.81      0.87       164
            Avocado       0.80      0.68      0.73       143
       Avocado ripe       0.89      0.98      0.93       166
             Banana       0.64      0.61      0.63       166
 Banana Lady Finger       0.62      0.67      0.64       152
         Banana Red       0.69      0.44      0.54       166
            Beetroot       0.44      0.42      0.43       150
          Blueberry       0.97      0.73      0.83       154
      Cabbage white       0.92      1.00      0.96        47
       Cactus fruit       0.88      0.87      0.87       166
         Cantaloupe       0.94      0.96      0.95       328
          Carambula       0.92      0.42      0.57       166
...
           accuracy                           0.81     23619
          macro avg       0.82      0.80      0.80     23619
       weighted avg       0.81      0.81      0.80     23619
```



Explanation:

- **The best hyperparameters were selected via a cross-validation process. The model achieved exceptionally high accuracy on the validation set (99.50%), but there is a significant drop in accuracy on the test set (80.57%). This suggests the model may have overfitted on the training data and struggled to generalize to unseen data.**
- **The precision, recall, and F1-scores vary across classes, with fruits like Apple showing strong performance, but classes like Apple Braeburn and Apple Red have lower scores, especially in precision and recall.**
- **The macro avg and weighted avg indicate that the model's performance is more balanced for some fruit classes but uneven for others.**

2. SVM

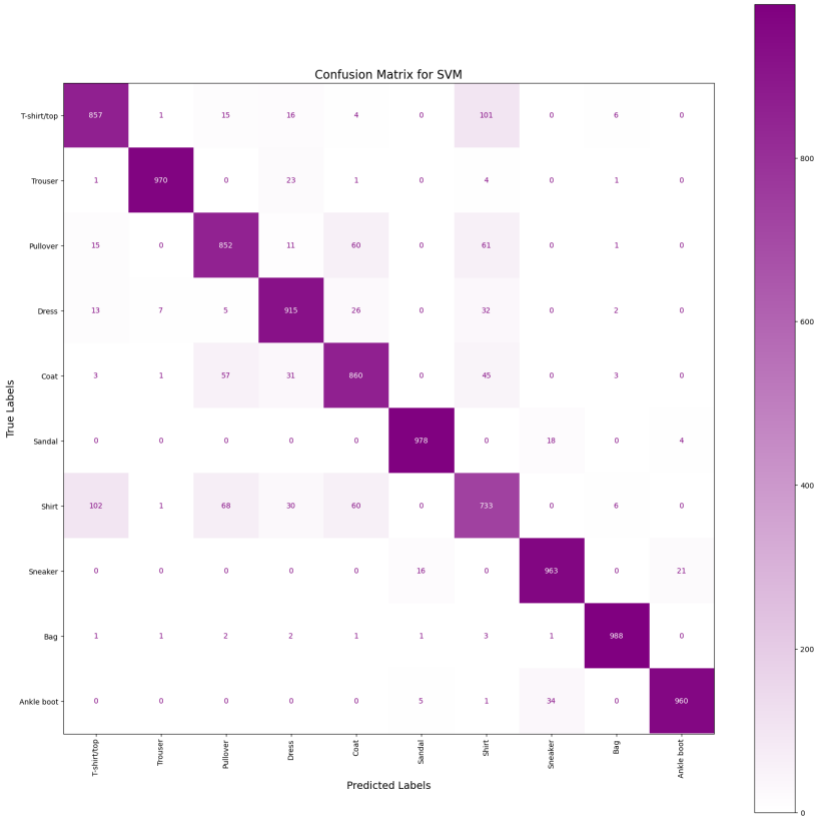
Fashion-MNIST Dataset

- **Best Hyperparameters:**
  - **C = 10, kernel = 'rbf'**
- **Accuracy:**
  - **Validation Set: 90.97%**
  - **Test Set: 90.76%**
- **Classification Report (Test Set):**

Test Classification Report for SVM:

	precision	recall	f1-score	support
T-shirt/top	0.86	0.86	0.86	1000
Trouser	0.99	0.97	0.98	1000
Pullover	0.85	0.85	0.85	1000
Dress	0.89	0.92	0.90	1000
Coat	0.85	0.86	0.85	1000
Sandal	0.98	0.98	0.98	1000
Shirt	0.75	0.73	0.74	1000
Sneaker	0.95	0.96	0.96	1000
Bag	0.98	0.99	0.98	1000
Ankle boot	0.97	0.96	0.97	1000
accuracy			0.91	10000
macro avg	0.91	0.91	0.91	10000
weighted avg	0.91	0.91	0.91	10000

- **Confusion Matrix**



Explanation:

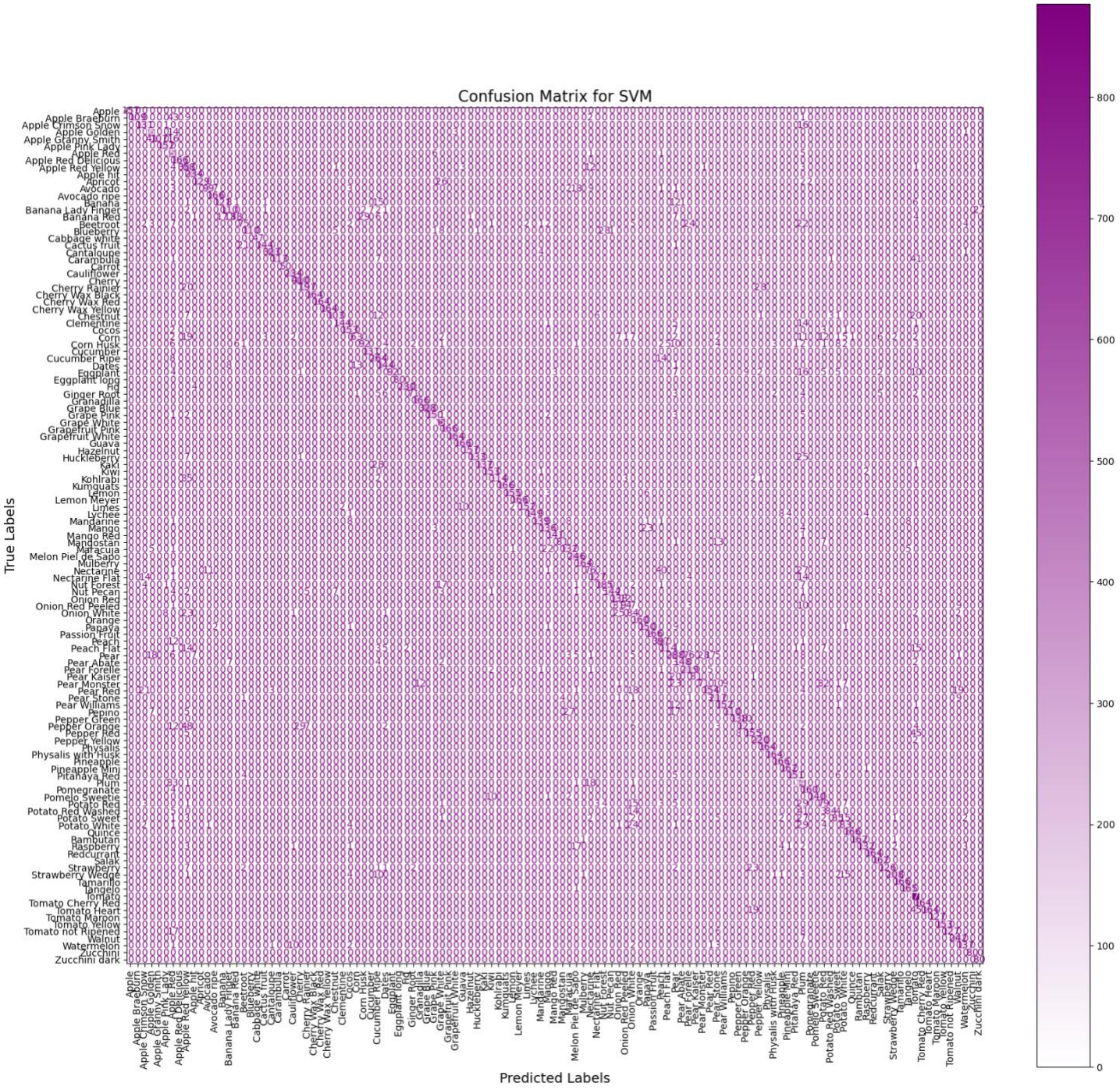
- The **best hyperparameters** were selected via **cross-validation**, with the **SVM model** showing **excellent performance** on both the validation and test sets, achieving **90.76% accuracy** on the test set.
- **Precision**, **recall**, and **F1-scores** for most classes are **high**, with **Trouser** and **Sandal** performing especially well, but **Shirt** has relatively lower values in all metrics (precision 0.75, recall 0.73, F1-score 0.74).
- The **macro avg** and **weighted avg** confirm the model's overall balanced performance across all classes.

Fruits-360 Dataset

- **Best Hyperparameters:**
  - **C = 1, kernel = 'rbf'**
- **Accuracy:**
  - **Validation Set: 99.50%**
  - **Test Set: 78.00%**
- **Classification Report (Test Set):**

```
Test Classification Report for SVM:
                     precision    recall  f1-score   support

              Apple       1.00      1.00      1.00       157
     Apple Braeburn       1.00      0.66      0.80       164
 Apple Crimson Snow       0.74      0.89      0.81       148
       Apple Golden       0.86      0.96      0.91       485
 Apple Granny Smith       1.00      0.65      0.79       164
    Apple Pink Lady       0.94      1.00      0.97       152
          Apple Red       0.68      0.97      0.80       472
Apple Red Delicious       1.00      1.00      1.00       166
   Apple Red Yellow       0.62      0.93      0.75       383
          Apple hit       0.95      1.00      0.97       234
            Apricot       1.00      0.79      0.88       164
            Avocado       0.89      0.69      0.78       143
       Avocado ripe       0.95      1.00      0.97       166
             Banana       0.88      0.73      0.80       166
 Banana Lady Finger       0.75      0.72      0.74       152
         Banana Red       0.93      0.53      0.67       166
           Beetroot       0.72      0.50      0.59       150
          Blueberry       1.00      0.71      0.83       154
      Cabbage white       1.00      1.00      1.00        47
       Cactus fruit       0.97      0.87      0.91       166
         Cantaloupe       0.99      0.98      0.99       328
          Carambula       1.00      0.68      0.81       166
...
           accuracy                           0.87     23619
          macro avg       0.91      0.86      0.87     23619
       weighted avg       0.89      0.87      0.87     23619
```



- **Explanation:**
  - **The best hyperparameters were selected via cross-validation. The model showed very high accuracy on the validation set (99.50%), but there was a significant drop in performance on the test set (78% accuracy). This indicates possible overfitting, where the model learned well on the validation data but struggled to generalize on the test data.**
  - **Precision, recall, and F1-scores for each class show variability, with fruits like Apple performing well (precision 0.80, recall 1.00), while other classes like Apple Braeburn and Apple Red have lower metrics.**
  - **The macro avg and weighted avg reflect the model's overall performance decline on the test set, likely due to imbalanced class distribution or model overfitting.**

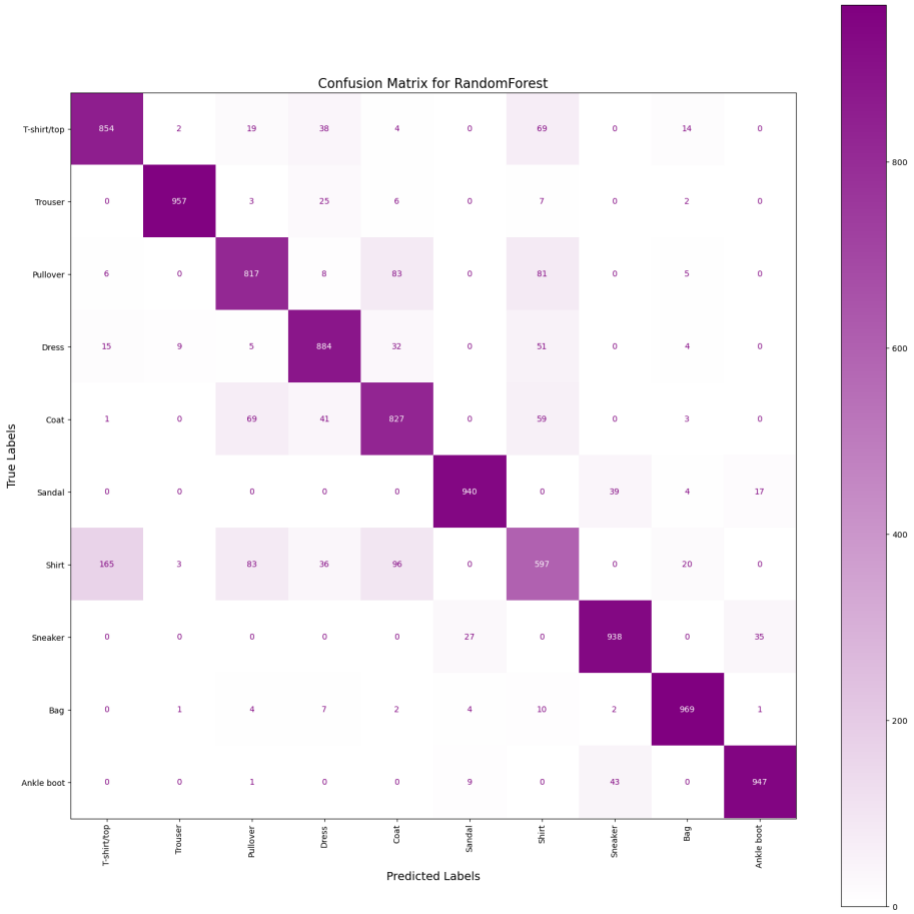
3. **RandomForest**

Fashion-MNIST Dataset

- **Best Hyperparameters:**
  - **max_depth = 20, max_samples = 1.0, n_estimators = 200**
- **Accuracy:**
  - **Validation Set: 87.32%**
  - **Test Set: 87.30%**
- **Classification Report (Test Set):**

Test Classification Report for RandomForest:

	precision	recall	f1-score	support
T-shirt/top	0.82	0.85	0.84	1000
Trouser	0.98	0.96	0.97	1000
Pullover	0.82	0.82	0.82	1000
Dress	0.85	0.88	0.87	1000
Coat	0.79	0.83	0.81	1000
Sandal	0.96	0.94	0.95	1000
Shirt	0.68	0.60	0.64	1000
Sneaker	0.92	0.94	0.93	1000
Bag	0.95	0.97	0.96	1000
Ankle boot	0.95	0.95	0.95	1000
accuracy			0.87	10000
macro avg	0.87	0.87	0.87	10000
weighted avg	0.87	0.87	0.87	10000



Explanation:

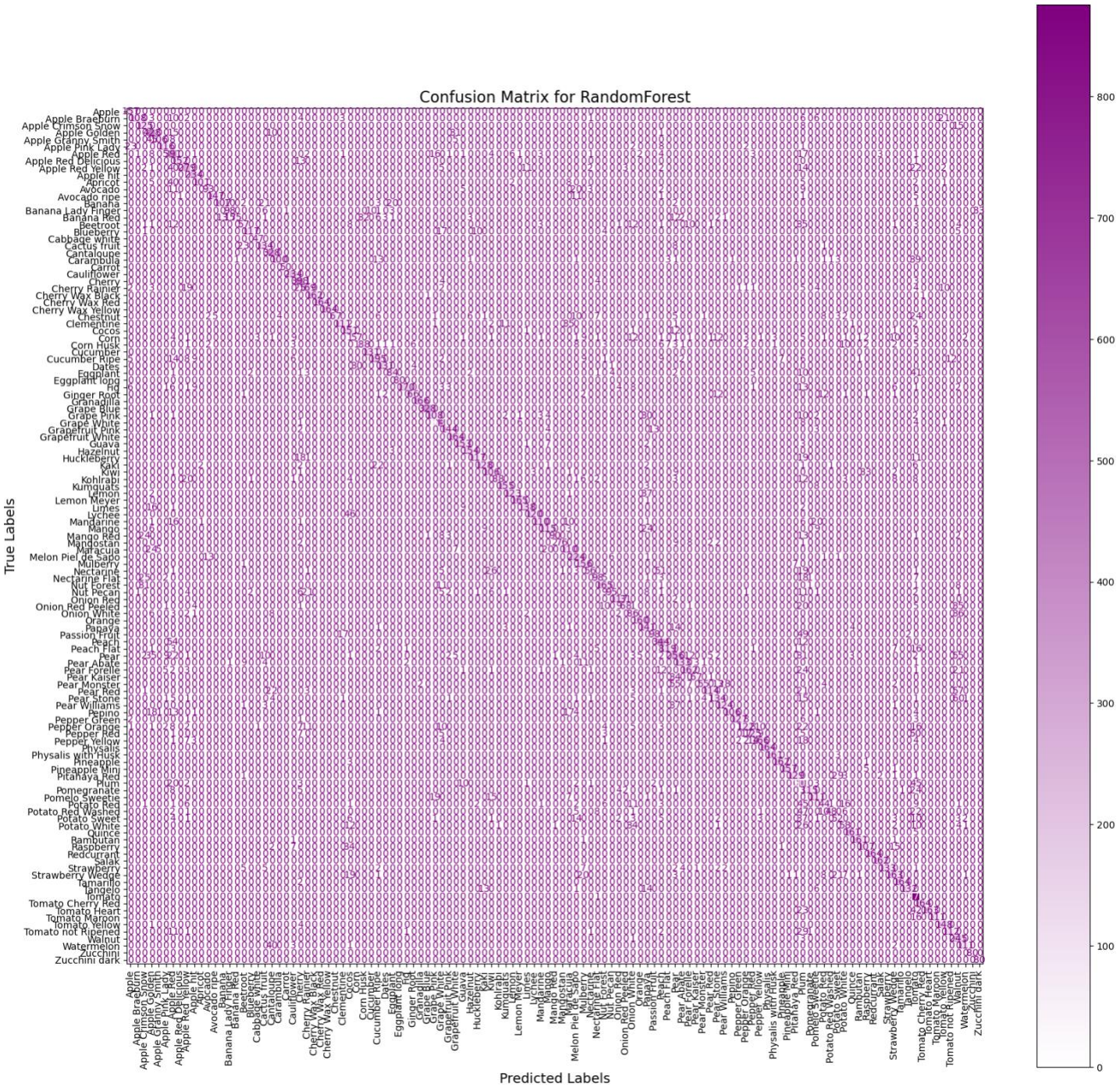
- **The best hyperparameters were selected using cross-validation. The Random Forest model performed with 87.30% accuracy on the test set, indicating strong performance, with a relatively balanced precision and recall across most classes.**
- **Precision, recall, and F1-scores are good for most classes. For example, Trouser and Sandal show high scores (close to 0.97), but Shirt has a lower precision (0.68) and recall (0.60), which suggests it is more challenging for the model to classify accurately.**
- **The macro avg and weighted avg indicate that the model performs well across all classes, but the performance is slightly lower for less frequent or more challenging classes like Shirt.**

Fruits-360 Dataset

- **Best Hyperparameters:**
  - **max_depth = 20, max_samples = 1.0, n_estimators = 200**
- **Accuracy:**
  - **Validation Set: 99.50%**
  - **Test Set: 78.02%**
- **Classification Report (Test Set):**

Test Classification Report for RandomForest:

	precision	recall	f1-score	support
Apple	0.80	1.00	0.89	157
Apple Braeburn	0.98	0.66	0.79	164
Apple Crimson Snow	0.59	0.84	0.69	148
Apple Golden	0.71	0.88	0.78	485
Apple Granny Smith	0.95	0.65	0.77	164
Apple Pink Lady	0.82	0.76	0.79	152
Apple Red	0.54	0.83	0.65	472
Apple Red Delicious	0.99	0.92	0.95	166
Apple Red Yellow	0.77	0.73	0.75	383
Apple hit	0.89	1.00	0.94	234
Apricot	0.97	0.62	0.75	164
Avocado	0.85	0.65	0.74	143
Avocado ripe	0.97	0.89	0.92	166
Banana	0.89	0.64	0.75	166
Banana Lady Finger	0.79	0.64	0.71	152
Banana Red	0.99	0.45	0.62	166
Beetroot	0.57	0.38	0.46	150
Blueberry	0.96	0.76	0.85	154
Cabbage white	0.92	1.00	0.96	47
Cactus fruit	0.73	0.81	0.77	166
Cantaloupe	0.79	1.00	0.88	328
Carambula	0.96	0.60	0.74	166
...				
accuracy			0.78	23619
macro avg	0.84	0.77	0.79	23619
weighted avg	0.81	0.78	0.78	23619



Explanation:

The best hyperparameters for the Random Forest model were selected using GridSearch with cross-validation. While the model achieved 99.50% accuracy on the validation set, it experienced a significant drop in performance on the test set (78.02% accuracy). This suggests that the model overfitted to the validation data, learning patterns specific to the training set, but struggled to generalize to the test set.

The precision, recall, and F1-scores for certain classes, such as Apple (precision: 0.80, recall: 1.00) and Apple Red Delicious (precision: 0.99, recall: 0.92), were high, indicating strong performance for these fruits. However, other classes like Apple Braeburn and Beetroot showed lower performance metrics, which suggests that the model struggled to distinguish these less frequent or more challenging fruits.

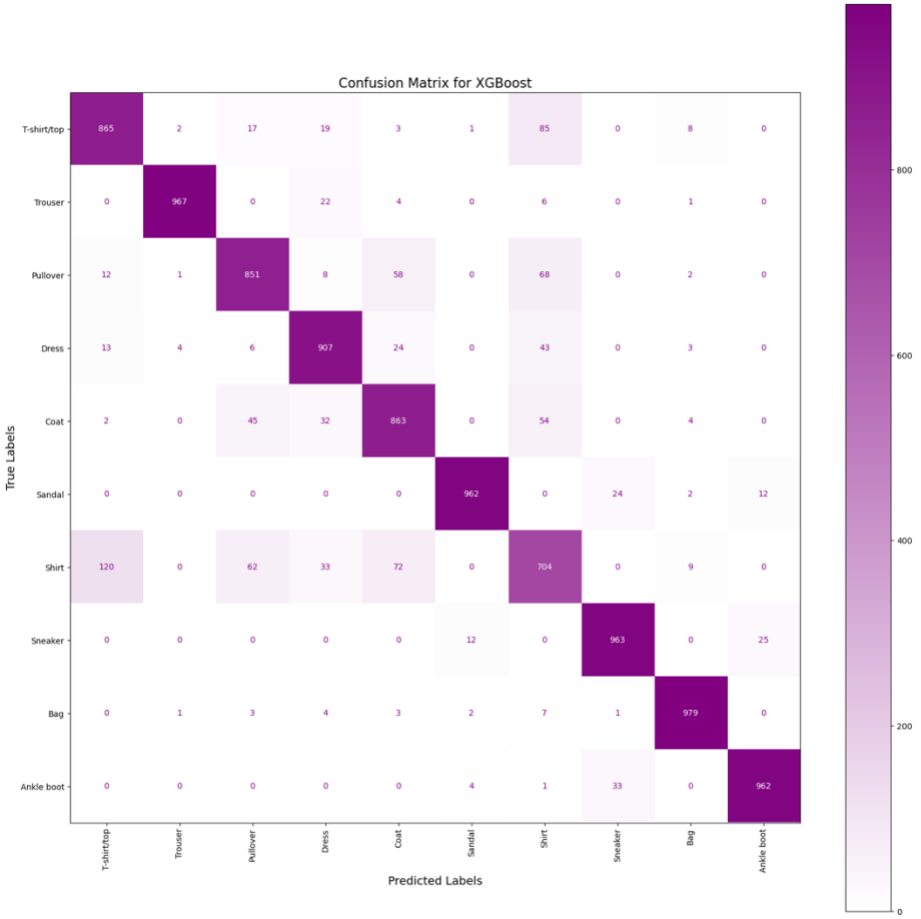
4. Gradient Boosted Trees (XGBoost) Evaluation

Fashion-MNIST Dataset

- **Best Hyperparameters:**
  - **learning_rate = 0.2, max_depth = 7, n_estimators = 200**
- **Accuracy:**
  - **Validation Set: 90.43%**
  - **Test Set: 90.23%**
- **Classification Report (Test Set):**

Test Classification Report for XGBoost:

	precision	recall	f1-score	support
T-shirt/top	0.85	0.86	0.86	1000
Trouser	0.99	0.97	0.98	1000
Pullover	0.86	0.85	0.86	1000
Dress	0.88	0.91	0.90	1000
Coat	0.84	0.86	0.85	1000
Sandal	0.98	0.96	0.97	1000
Shirt	0.73	0.70	0.72	1000
Sneaker	0.94	0.96	0.95	1000
Bag	0.97	0.98	0.98	1000
Ankle boot	0.96	0.96	0.96	1000
accuracy			0.90	10000
macro avg	0.90	0.90	0.90	10000
weighted avg	0.90	0.90	0.90	10000



Explanation:

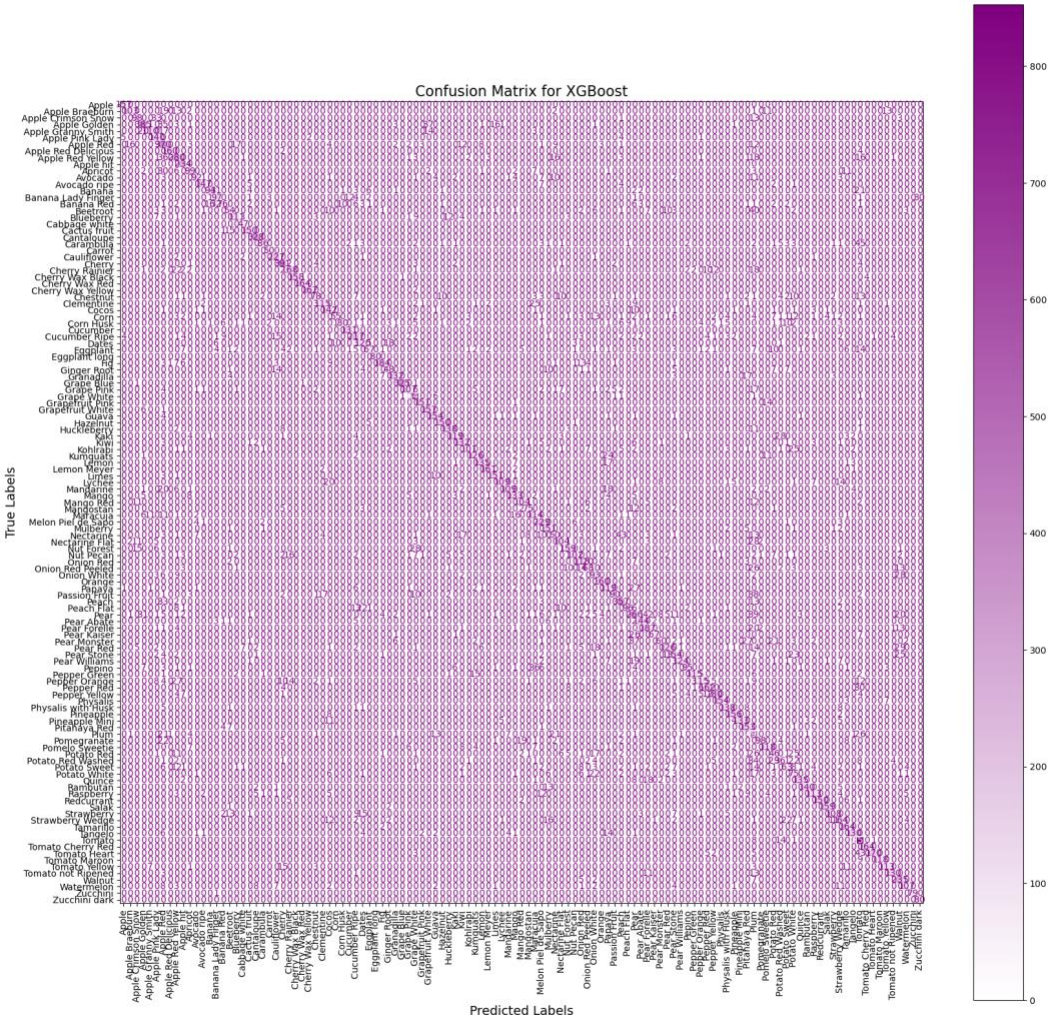
- The **best hyperparameters** were selected using **GridSearch** with **cross-validation**. The **XGBoost model** achieved **90.23% accuracy** on the test set, demonstrating strong classification performance across most classes.
- **Precision**, **recall**, and **F1-scores** are high for most classes, particularly **Trouser** and **Sandal** (precision: 0.99, recall: 0.97, F1-score: 0.98), while **Shirt** had lower values (precision: 0.73, recall: 0.70, F1-score: 0.72), indicating that it is a more challenging class for the model.
- The **macro avg** and **weighted avg** suggest that the model is well-balanced overall, with relatively consistent performance across classes.

Fruits-360 Dataset

- **Best Hyperparameters**:
  - learning_rate = 0.2, max_depth = 3, n_estimators = 200
- **Accuracy**:
  - **Validation Set**: 99.37%
  - **Test Set**: 77.02%
- **Classification Report** (Test Set):

Test Classification Report for XGBoost:

	precision	recall	f1-score	support
Apple	0.94	1.00	0.97	157
Apple Braeburn	0.84	0.63	0.72	164
Apple Crimson Snow	0.70	0.66	0.68	148
Apple Golden	0.83	0.79	0.81	485
Apple Granny Smith	0.85	0.67	0.75	164
Apple Pink Lady	0.63	0.92	0.75	152
Apple Red	0.53	0.78	0.63	472
Apple Red Delicious	0.99	0.96	0.98	166
Apple Red Yellow	0.63	0.73	0.68	383
Apple hit	0.85	1.00	0.92	234
Apricot	0.72	0.60	0.66	164
Avocado	0.83	0.64	0.72	143
Avocado ripe	0.93	0.89	0.91	166
Banana	0.83	0.57	0.67	166
Banana Lady Finger	0.55	0.64	0.59	152
Banana Red	0.77	0.46	0.57	166
Beetroot	0.53	0.36	0.43	150
Blueberry	0.80	0.72	0.76	154
Cabbage white	0.76	1.00	0.86	47
Cactus fruit	0.89	0.90	0.90	166
Cantaloupe	0.94	1.00	0.97	328
Carambula	0.91	0.52	0.66	166
...				
accuracy			0.77	23619
macro avg	0.80	0.76	0.77	23619
weighted avg	0.79	0.77	0.77	23619



CONCLUSIONS

Fashion-MNIST Dataset:

- **Best Models**: **SVM** and **XGBoost** performed the best, with **SVM** achieving **90.76%** test accuracy and **XGBoost** close behind at **90.23%**. Both models showed strong performance across most classes, especially **Trouser** and **Sandal**.
- **Conclusion**: **SVM** and **XGBoost** are the top performers for **Fashion-MNIST**, with **SVM** slightly ahead in test accuracy.

Fruits-360 Dataset:

- **Best Models**: All models (Logistic Regression, SVM, Random Forest, XGBoost) had **excellent validation accuracy** (over 99%) but faced **significant overfitting** on the test set (around **77-78%** accuracy).
- **Conclusion**: **Logistic Regression** and **Random Forest** performed similarly on the test set (**78%** accuracy), while **XGBoost** slightly lagged at **77.02%**. Overfitting is an issue, and addressing it could improve generalization.