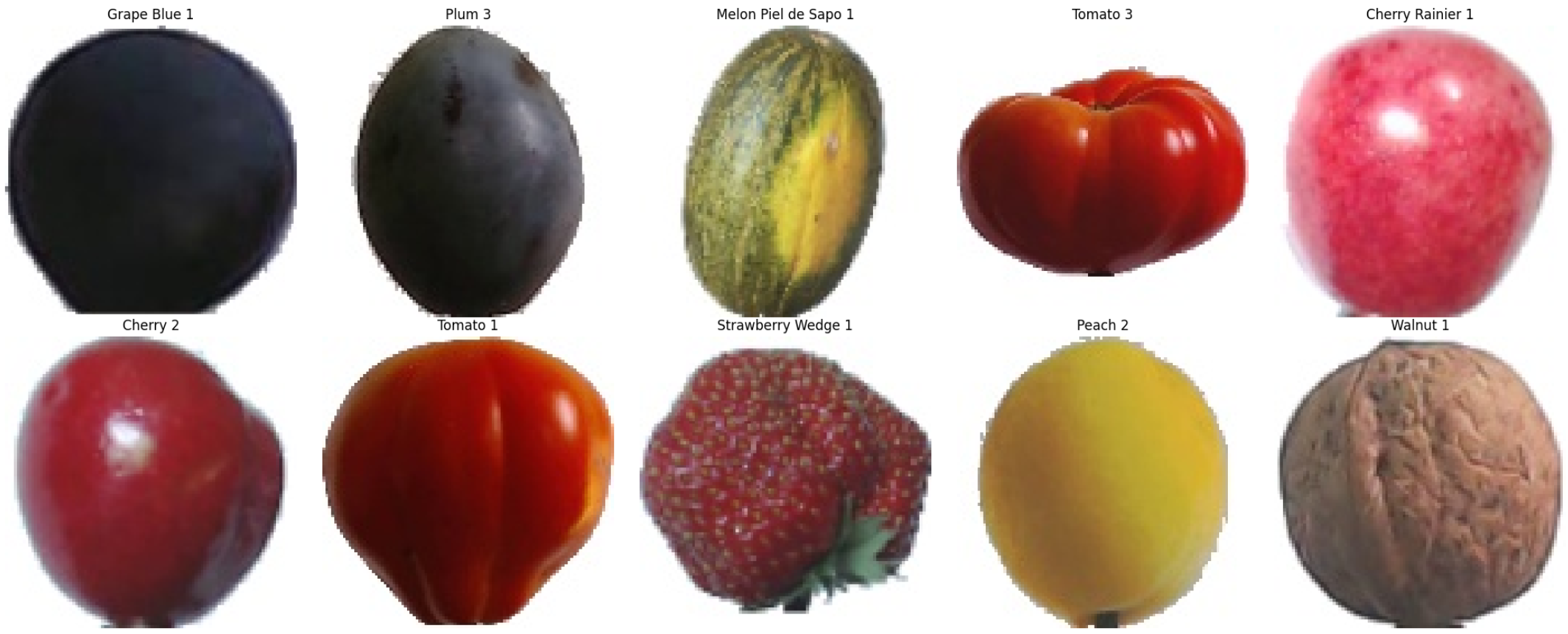
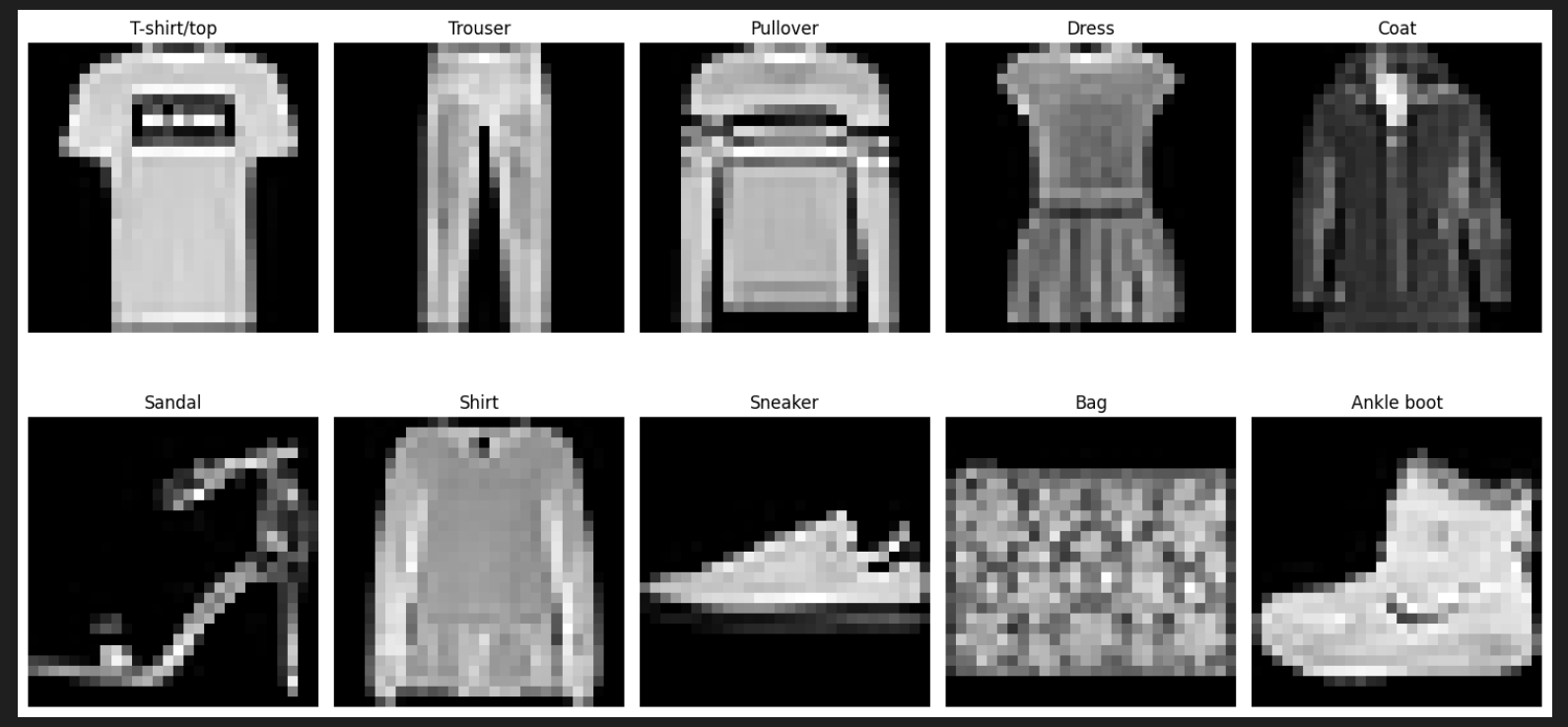
Tema 1 ML

De Florea Radu

# Descrierea fluxului propus pentru extragerea atributelor

4.1

La primul task, am folosit doua metode pentru extragerea de atribute : PCA (Principal Component Analysis) si HOG (Histogram of Oriented Gradients). Aceaste filtre permit captarea informațiilor globale (prin PCA) și locale (prin HOG).

4.2

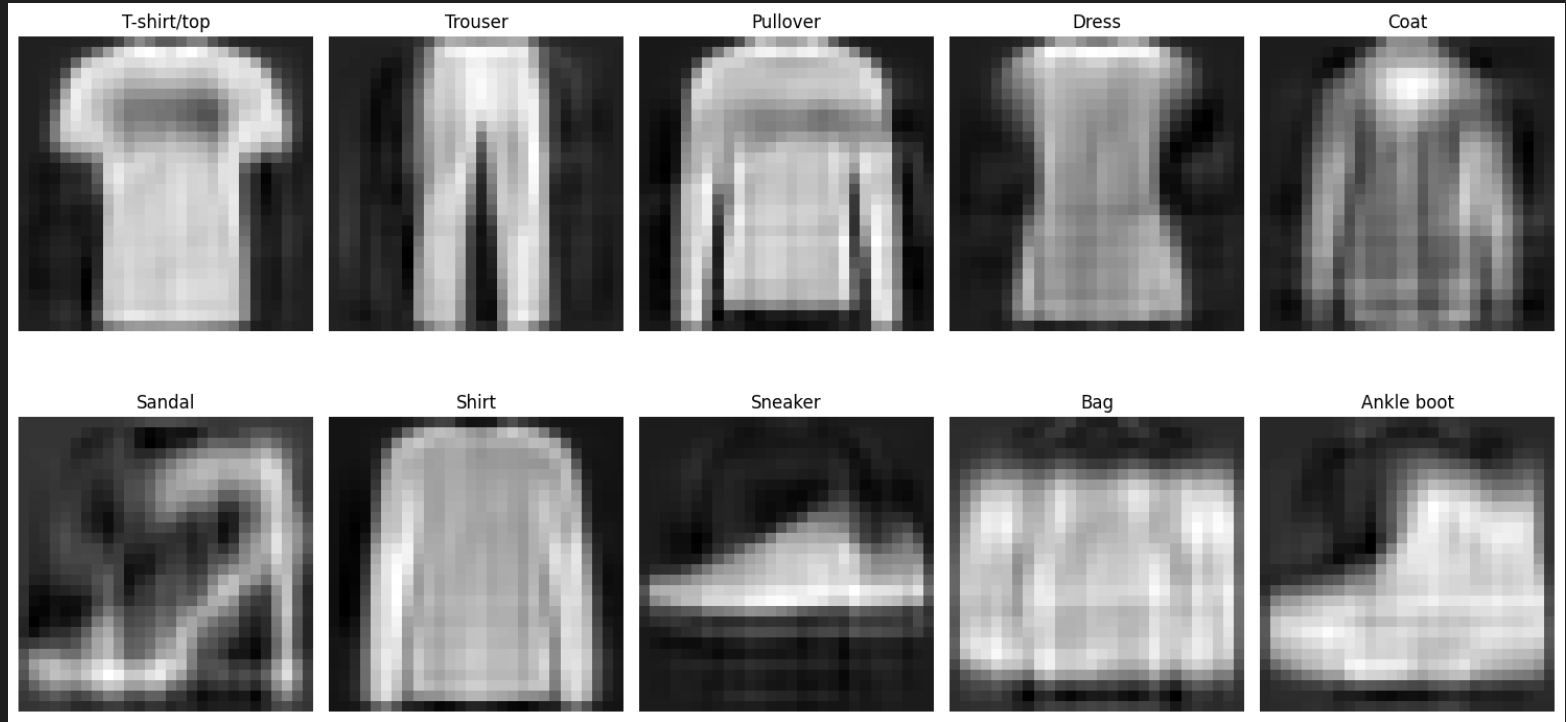
PCA

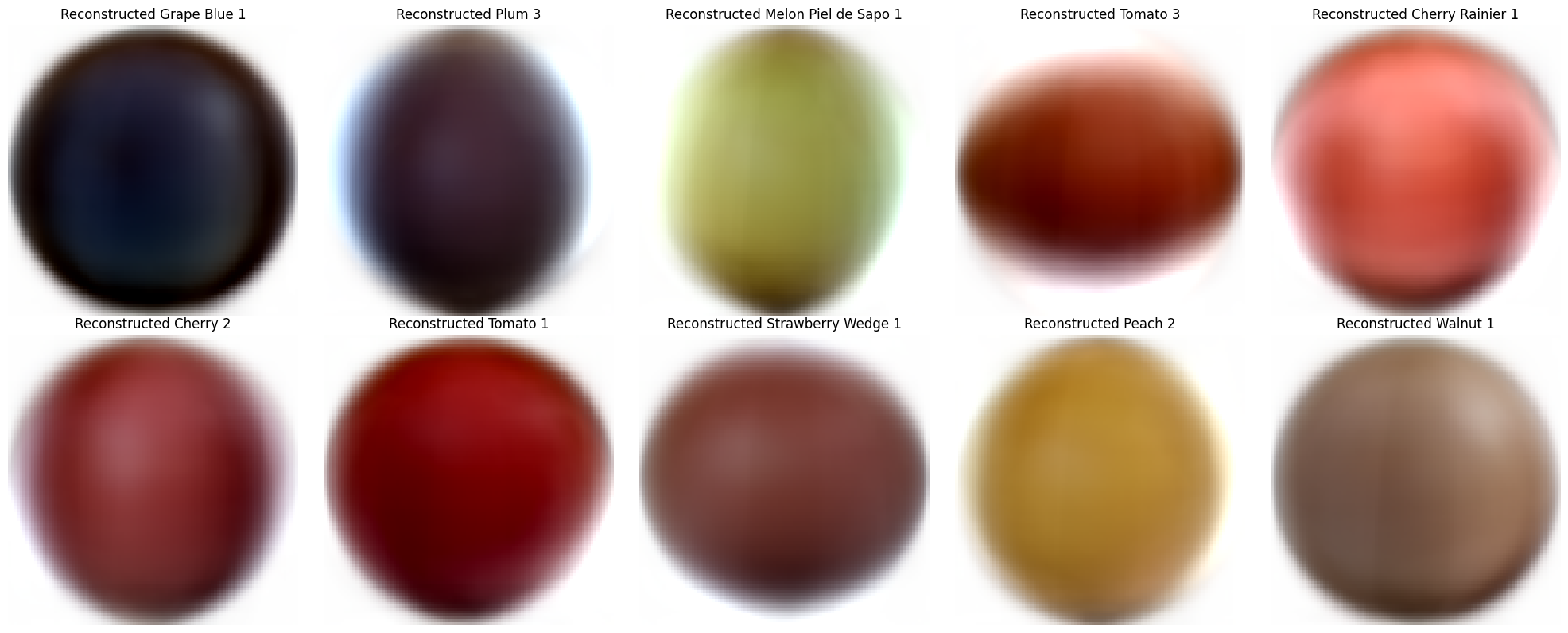
PCA reduce dimensionalitatea imaginilor și păstrează variația importantă.

Mai întâi, imaginile au fost **aplatizate** (transformate într-un vector unidimensional) pentru a obține o reprezentare potrivită metodei PCA. Această operație a transformat fiecare imagine într-un vector de dimensiune 100x100x3 (pentru imagini RGB).

Ulterior, am utilizat metoda **fit\_transform** din PCA pentru a ajusta modelul PCA pe datele de intrare și a reduce dimensionalitatea acestora. Astfel, PCA a identificat și selectat componentele principale care explică cea mai mare parte a variației din setul de date. Pentru a evalua performanța metodei PCA, am reconstruit 10 imagini din setul de train prin aplicarea inversării transformării PCA (**inverse\_transform**). Această operație permite vizualizarea informațiilor păstrate de componentele principale selectate în timpul reducerii dimensionalității.

A furnizat un set redus de caracteristici relevante, care descriu aspecte generale precum dominanța culorilor sau a texturilor. Se poate observa că varianta redusă a imaginii păstrează structura generală, dar elimină detaliile redundante. Pentru setul de date fashion\_mnist, am folosit 50 de componente principale, respectiv 10 pentru setul fruits360



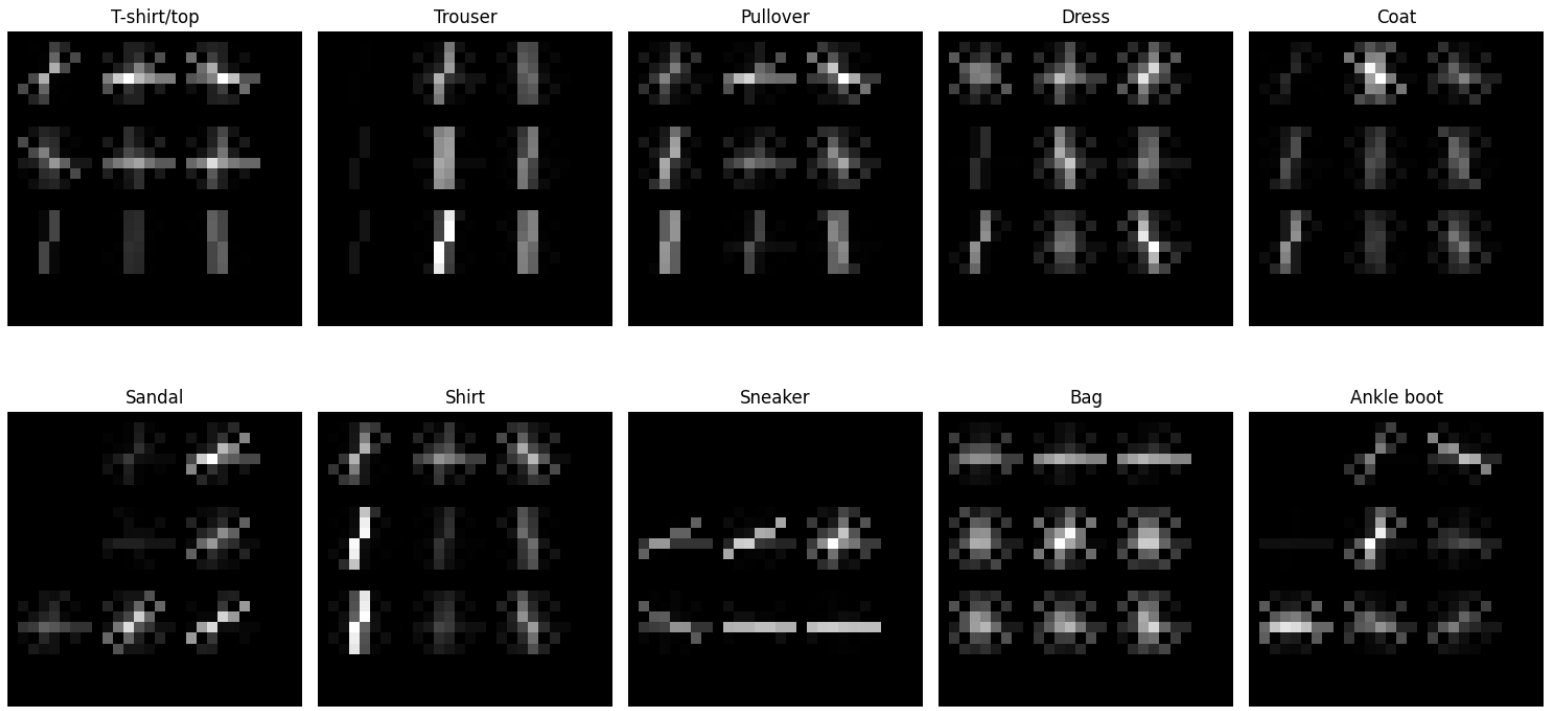


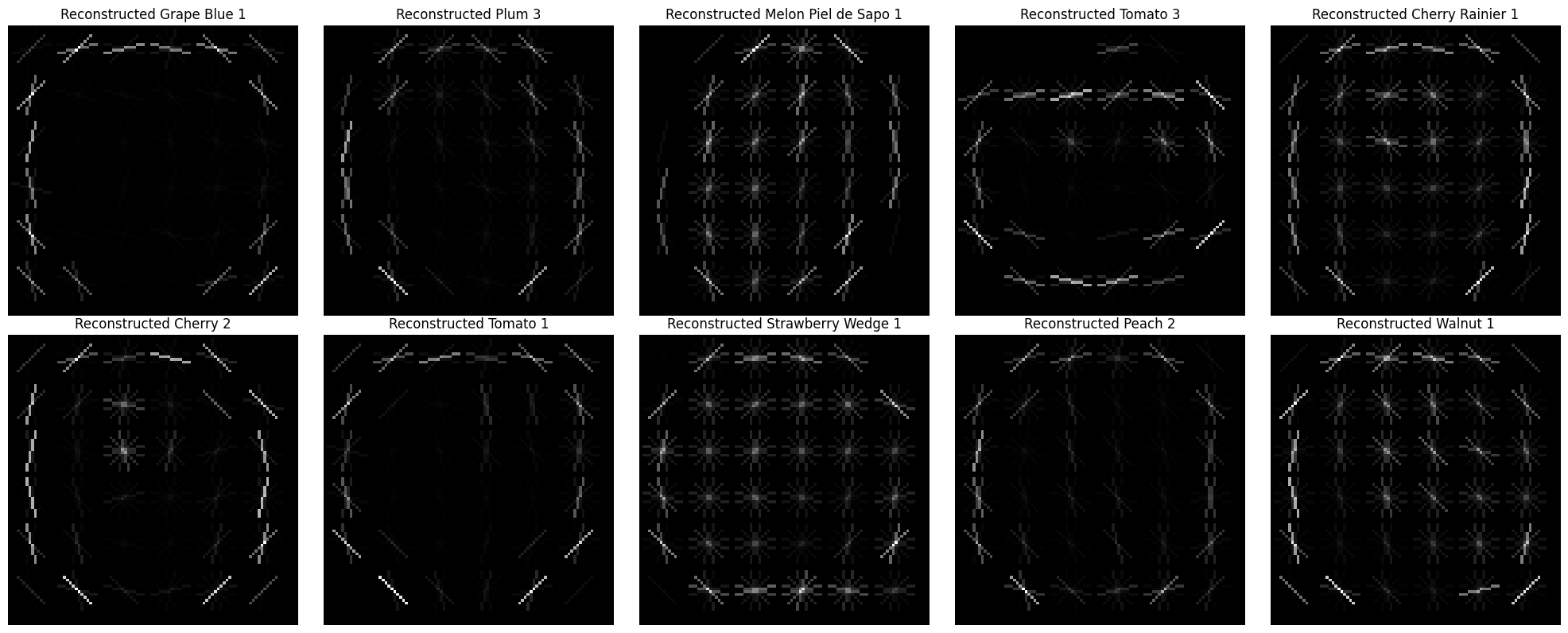
HOG

Pentru a analiza caracteristicile extrase cu metoda HOG, am selectat cele 10 imagini prezentate (pentru fiecare clasa) mai sus și am aplicat funcția **hog**. Imaginea a fost convertită în tonuri de gri, iar HOG a extras gradientele orientate utilizând 9 direcții, celule de dimensiune 8x8 și blocuri de 2x2 celule pentru setul de date fashion mnist, iar . gradientele orientate utilizând 6 direcții, celule de dimensiune 16x16 și blocuri de 1x1 celule pentru setul de date fruits360

Rezultatul vizualizează contururile și texturile identificate în imagine, demonstrând utilitatea HOG în evidențierea trăsăturilor locale relevante pentru clasificare.

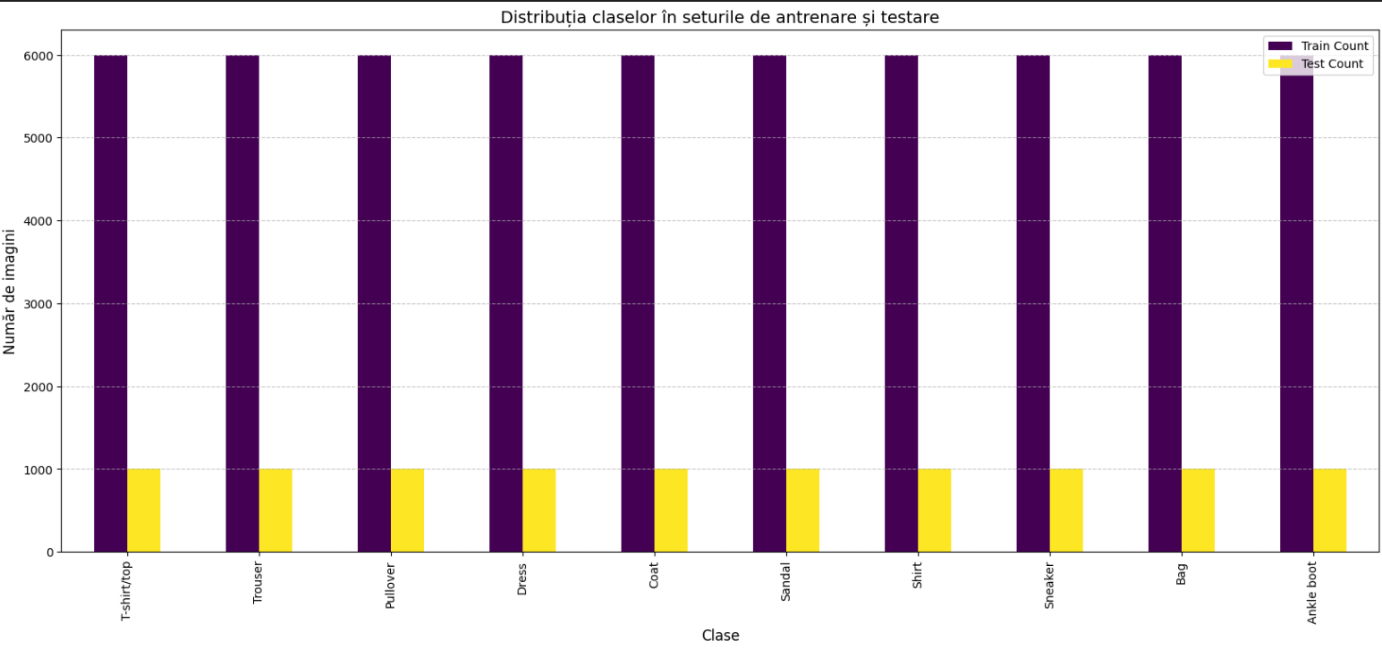
Gradientele captate includ atât contururile majore ale obiectului (cum ar fi marginea tomato 3), cât și detalii interne mai fine, esențiale pentru clasificarea obiectelor similare.





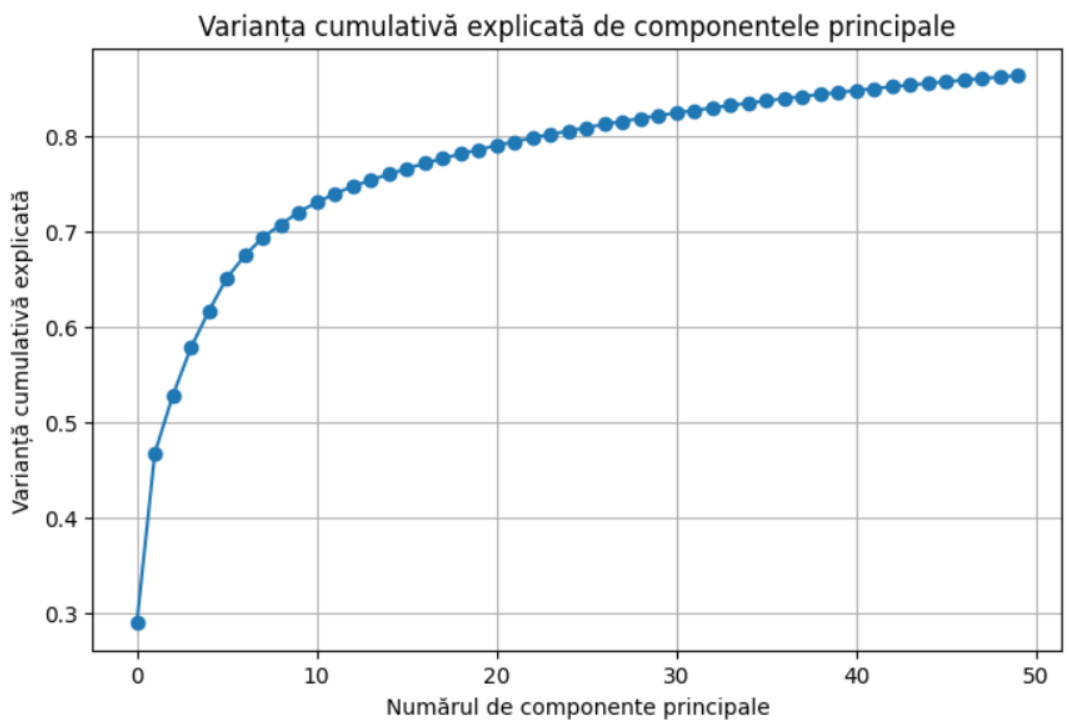
## Fashion\_mnist

Analiza echilibrului de clase



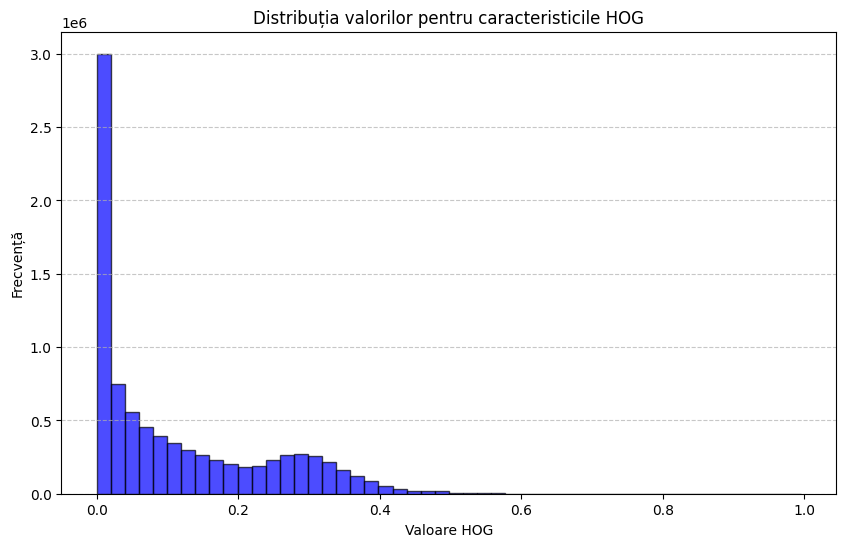
Întrucât datasetul este echilibrat, nu este necesar să aplic metode de echilibrare suplimentare, cum ar fi oversampling sau undersampling. Un dataset echilibrat asigură că modelul nu învață să favorizeze o clasă în detrimentul alteia.

PCA



Primele 10 componente explică peste 70% din varianță. Acest lucru indică faptul că datele au o redundanță ridicată, iar dimensiunea poate fi redusă semnificativ fără pierderea majorității informației.Graficul arată că aproximativ 30-40 de componente sunt necesare pentru a explica peste 85% din varianța totală. Acest prag este adesea utilizat ca o măsură practică pentru reducerea dimensionalității, păstrând totuși suficientă informație.

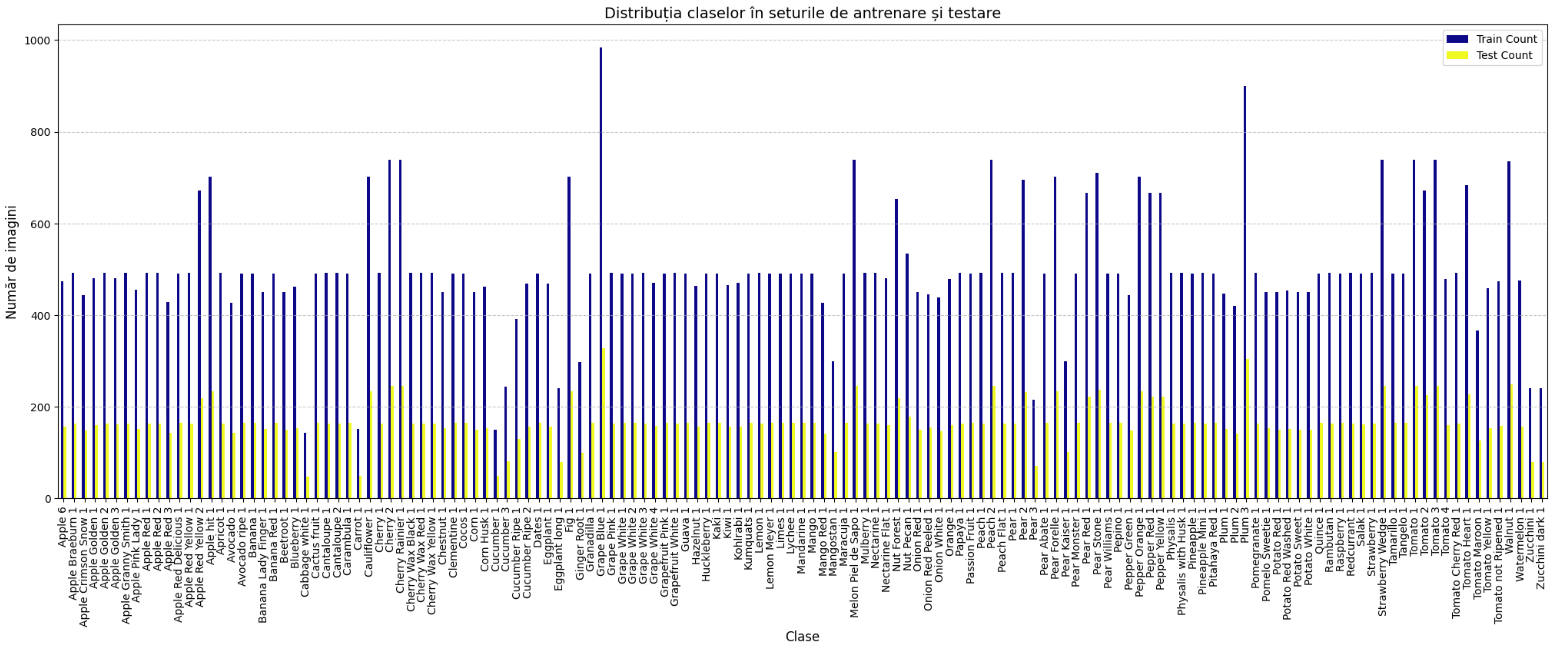
HOG - vizualizare cantitativa



Graficul ilustrează distribuția valorilor caracteristicilor HOG extrase din setul de date. Majoritatea valorilor sunt nule, ceea ce indică multe zone uniforme sau cu variații reduse în imagini. Totuși, există caracteristici cu valori mai mari, care corespund regiunilor cu gradienturi puternice, cum ar fi margini sau contururi pronunțate. Această distribuție subliniază importanța caracteristicilor cu valori mari, deoarece acestea oferă informații distinctive utile pentru clasificare.

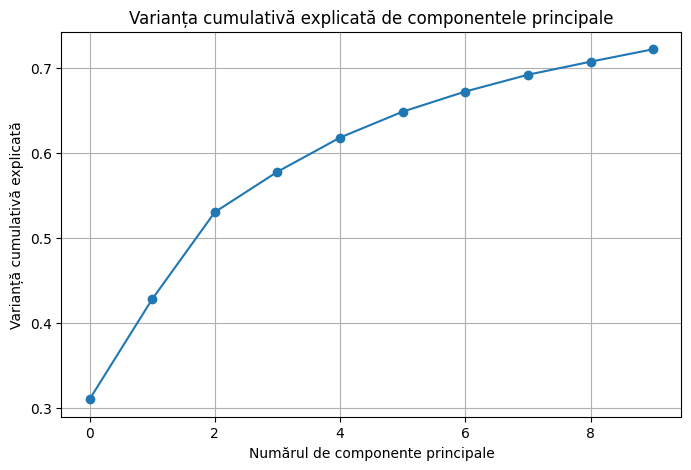
## Fruits360

Analiza echilibrului de clase



Distribuția claselor nu este uniformă Graficul indică un dezechilibru de clase. Din grafic, putem observa că distribuția claselor în setul de testare (barele galbene) urmează aceeași tendință generală ca în setul de antrenare (barele albastre). Acest lucru este pozitiv, deoarece asigură că modelul va fi testat pe date care reflectă distribuția pe care a fost antrenat.

PCA



Acest grafic arată cum varianța cumulativă explicată de componentele principale crește pe măsură ce adăugăm mai multe componente: Primele 2-3 componente explică cea mai mare parte din varianța, după a 4-a componentă, creșterea varianței explicate încetinește, la 7-9 componente, curba se aplatizează, aceste componente având un impact redus.

4.3

## Am ales Variance Threshold

Variance Threshold selectează atributele cu varianță suficient de mare, Nu ține cont de etichete, elimină doar atributele care sunt constante sau aproape constante.

Select Percentile selectează un procentaj din cele mai relevante atribute, este mai precis, dar mai lent. Din aceste cauze am ales Variance Threshold

**Dimensiuni înainte și după Variance Threshold pentru setul Fashion\_mnist:**

* Dimensiuni train înainte: (60000, 169), după: (60000, 95)
* Dimensiuni test înainte: (10000, 169), după: (10000, 95)

**Dimensiuni înainte și după Variance Threshold:**

* Dimensiuni train înainte: (70491, 226), după: (70491, 109)
* Dimensiuni test înainte: (23619, 226), după: (23619, 109)

## 4.4

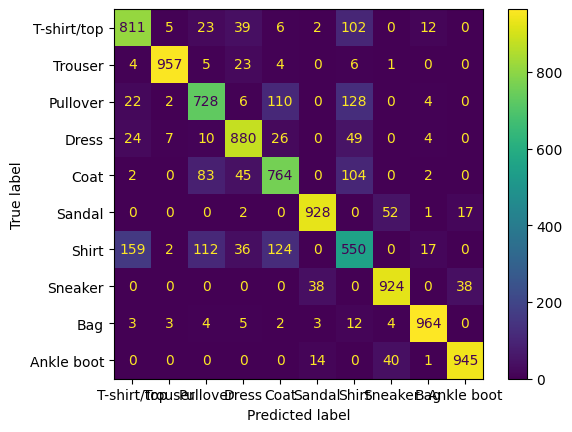
### Fashion\_­mnist

## LogisticRegression

Cei mai buni hiper-parametri: {'multi\_class': 'multinomial', 'C': 1}

Accuracy (Combined PCA + HOG): 0.8449

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-Score | Support |
| T-shirt/top | 0.79 | 0.81 | 0.80 | 1000 |
| Trouser | 0.98 | 0.96 | 0.97 | 1000 |
| Pullover | 0.75 | 0.73 | 0.74 | 1000 |
| Dress | 0.85 | 0.88 | 0.86 | 1000 |
| Coat | 0.74 | 0.76 | 0.75 | 1000 |
| Sandal | 0.94 | 0.93 | 0.94 | 1000 |
| Shirt | 0.58 | 0.55 | 0.56 | 1000 |
| Sneaker | 0.90 | 0.92 | 0.91 | 1000 |
| Bag | 0.96 | 0.96 | 0.96 | 1000 |
| Ankle boot | 0.94 | 0.94 | 0.94 | 1000 |
| Accuracy | - | - | - | 10000 |
| Macro avg | 0.84 | 0.85 | 0.84 | 10000 |
| Weighted avg | 0.84 | 0.85 | 0.84 | 10000 |

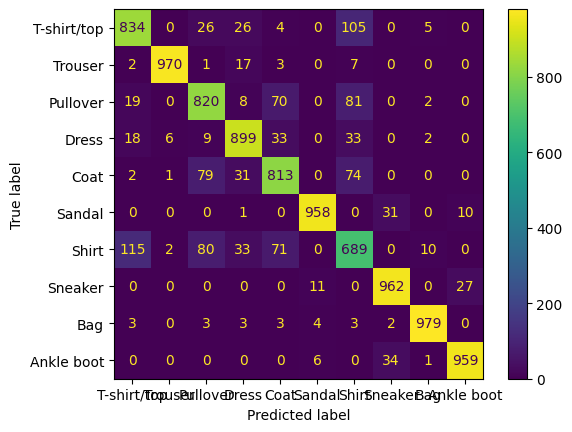


## SVM

Cei mai buni hiper-parametri SVM: {'kernel': 'rbf', 'C': 10}

Accuracy SVM (Combined PCA + HOG): 0.8883

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ****Class**** | ****Precision**** | ****Recall**** | ****F1-Score**** | ****Support**** |
| T-shirt/top | 0.84 | 0.83 | 0.84 | 1000 |
| Trouser | 0.99 | 0.97 | 0.98 | 1000 |
| Pullover | 0.81 | 0.82 | 0.81 | 1000 |
| Dress | 0.88 | 0.90 | 0.89 | 1000 |
| Coat | 0.82 | 0.81 | 0.81 | 1000 |
| Sandal | 0.98 | 0.96 | 0.97 | 1000 |
| Shirt | 0.69 | 0.69 | 0.69 | 1000 |
| Sneaker | 0.93 | 0.96 | 0.95 | 1000 |
| Bag | 0.98 | 0.98 | 0.98 | 1000 |
| Ankle boot | 0.96 | 0.96 | 0.96 | 1000 |
| **Accuracy** |  |  | **0.89** | **10000** |
| **Macro avg** | **0.89** | **0.89** | **0.89** | **10000** |
| **Weighted avg** | **0.89** | **0.89** | **0.89** | **10000** |
| **Class** | **Precision** | **Recall** | **F1-Score** | **Support** |

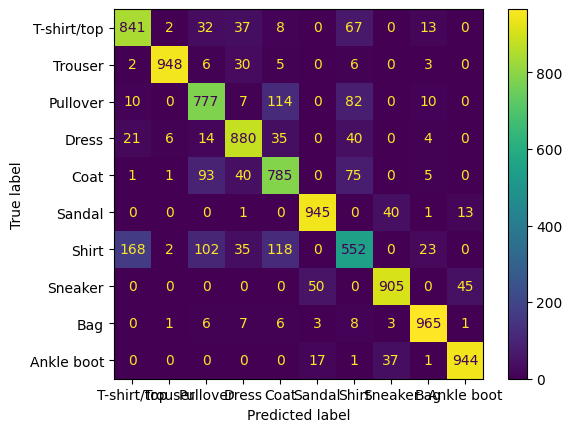


## RandomForestClassifier

Cei mai buni hiper-parametri Random Forest: {'n\_estimators': 200, 'max\_samples': 1.0, 'max\_depth': None}

Accuracy Random Forest (Combined PCA + HOG): 0.8542

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-Score | Support |
| T-shirt/top | 0.81 | 0.84 | 0.82 | 1000 |
| Trouser | 0.99 | 0.95 | 0.97 | 1000 |
| Pullover | 0.75 | 0.78 | 0.77 | 1000 |
| Dress | 0.85 | 0.87 | 0.86 | 1000 |
| Coat | 0.73 | 0.79 | 0.76 | 1000 |
| Sandal | 0.93 | 0.94 | 0.94 | 1000 |
| Shirt | 0.66 | 0.55 | 0.60 | 1000 |
| Sneaker | 0.92 | 0.91 | 0.91 | 1000 |
| Bag | 0.94 | 0.96 | 0.95 | 1000 |
| Ankle boot | 0.94 | 0.94 | 0.94 | 1000 |
| Accuracy |  |  | 0.85 | 10000 |
| Macro avg | 0.85 | 0.85 | 0.85 | 10000 |
| Weighted avg | 0.85 | 0.85 | 0.85 | 10000 |
| Class | Precision | Recall | F1-Score | Support |



### GradientBoostedTrees -- Xgboost

Cei mai buni hiper-parametri XGBClassifier: {'subsample': 0.75, 'n\_estimators': 200, 'max\_depth': 6, 'learning\_rate': 0.2, 'colsample\_bytree': 1.0}

Accuracy XGBClassifier: 0.8818

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-Score | Support |
| T-shirt/top | 0.84 | 0.84 | 0.84 | 1000 |
| Trouser | 0.99 | 0.96 | 0.98 | 1000 |
| Pullover | 0.80 | 0.80 | 0.80 | 1000 |
| Dress | 0.88 | 0.89 | 0.89 | 1000 |
| Coat | 0.78 | 0.83 | 0.81 | 1000 |
| Sandal | 0.97 | 0.96 | 0.96 | 1000 |
| Shirt | 0.68 | 0.65 | 0.67 | 1000 |
| Sneaker | 0.93 | 0.95 | 0.94 | 1000 |
| Bag | 0.98 | 0.97 | 0.97 | 1000 |
| Ankle boot | 0.96 | 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 |

### A purple and yellow squares with white text Description automatically generated

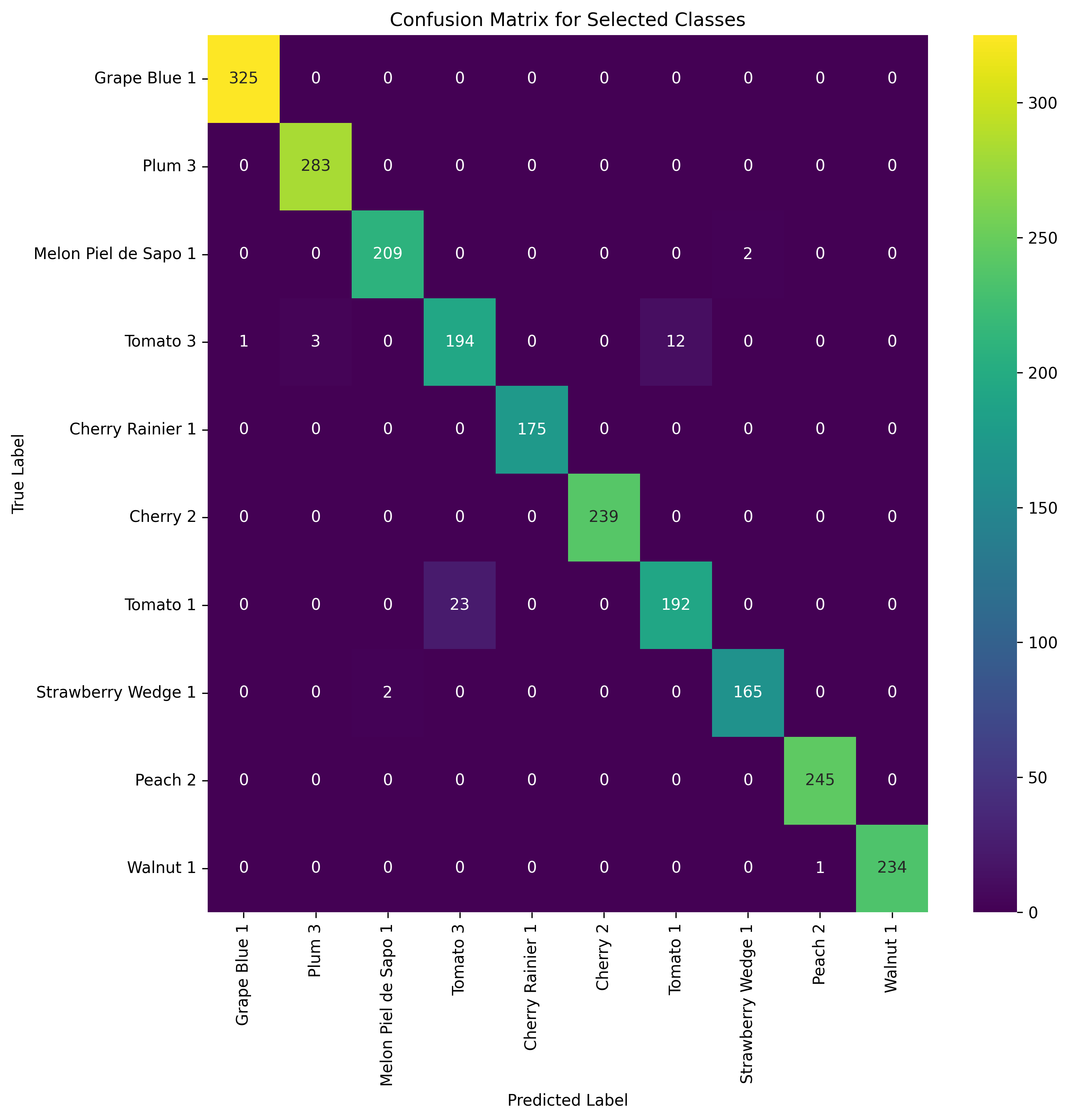
### Fruits360

## LogisticRegression

Cei mai buni hiper-parametri: {'multi\_class': 'multinomial', 'C': 10}

Accuracy (Combined PCA + HOG): 0.7909733688979211

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-Score | Support |
| Grape Blue 1 | 0.92 | 0.99 | 0.95 | 328 |
| Plum 3 | 0.89 | 0.93 | 0.91 | 304 |
| Melon Piel de Sapo 1 | 0.78 | 0.85 | 0.81 | 246 |
| Tomato 3 | 0.74 | 0.79 | 0.76 | 246 |
| Cherry Rainier 1 | 0.87 | 0.71 | 0.78 | 246 |
| Cherry 2 | 0.92 | 0.97 | 0.95 | 246 |
| Tomato 1 | 0.90 | 0.78 | 0.84 | 246 |
| Strawberry Wedge 1 | 0.53 | 0.67 | 0.59 | 246 |
| Peach 2 | 0.98 | 1.00 | 0.99 | 246 |
| Walnut 1 | 0.82 | 0.94 | 0.88 | 249 |
| Accuracy |  |  | 0.79 | 23619 |
| Macro Avg | 0.80 | 0.79 | 0.79 | 23619 |
| Weighted Avg | 0.80 | 0.79 | 0.79 | 23619 |

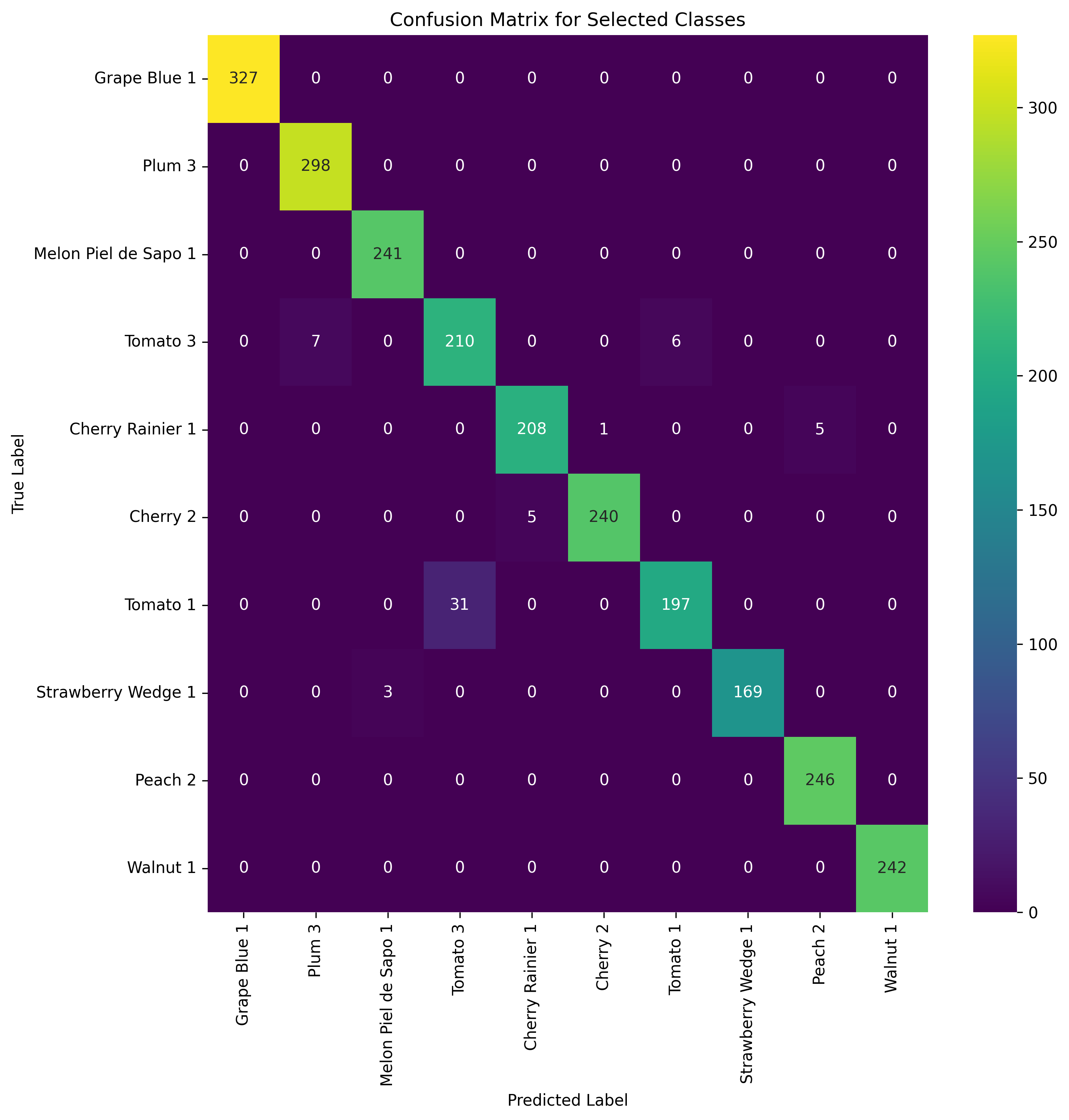


## SVM

Cei mai buni hiper-parametri SVM: {'kernel': 'linear', 'C': 1}

Accuracy SVM (Combined PCA + HOG): 0.8573605995173378

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-Score | Support |
| Grape Blue 1 | 0.98 | 1.00 | 0.99 | 328 |
| Plum 3 | 0.94 | 0.98 | 0.96 | 304 |
| Melon Piel de Sapo 1 | 0.89 | 0.98 | 0.93 | 246 |
| Tomato 3 | 0.73 | 0.85 | 0.79 | 246 |
| Cherry Rainier 1 | 0.88 | 0.85 | 0.86 | 246 |
| Cherry 2 | 0.97 | 0.98 | 0.97 | 246 |
| Tomato 1 | 0.94 | 0.80 | 0.86 | 246 |
| Strawberry Wedge 1 | 0.77 | 0.69 | 0.73 | 246 |
| Peach 2 | 0.98 | 1.00 | 0.99 | 246 |
| Walnut 1 | 0.85 | 0.97 | 0.91 | 249 |
| Accuracy |  |  | 0.86 | 23619 |
| Macro Avg | 0.86 | 0.86 | 0.85 | 23619 |
| Weighted Avg | 0.86 | 0.86 | 0.85 | 23619 |

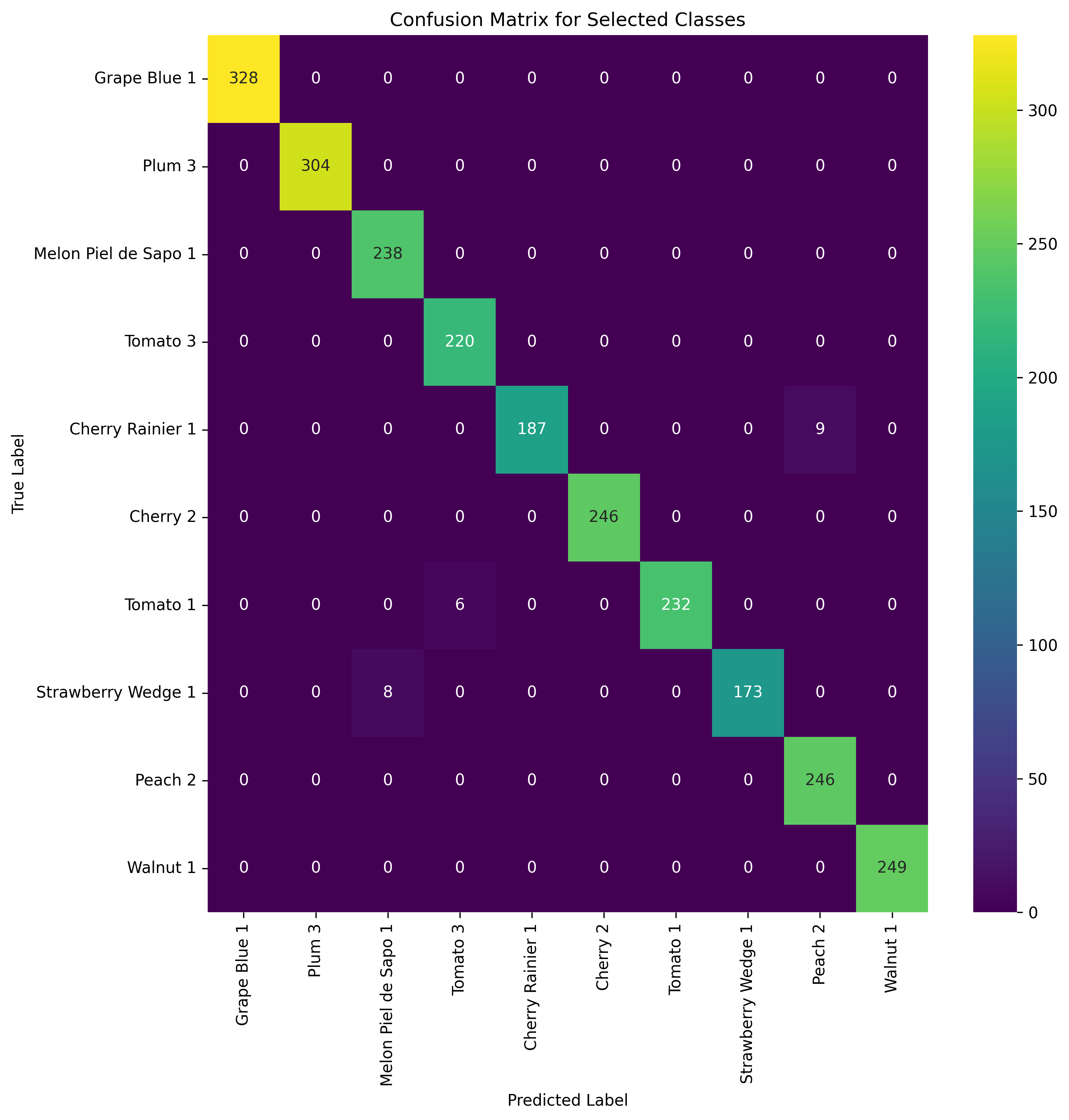


## RandomForestClassifier

Cei mai buni hiper-parametri Random Forest: {'n\_estimators': 200, 'max\_samples': 1.0, 'max\_depth': None}

Accuracy Random Forest (Combined PCA + HOG): 0.8693001397180237

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-Score | Support |
| Grape Blue 1 | 0.98 | 1.00 | 0.99 | 328 |
| Plum 3 | 0.72 | 1.00 | 0.84 | 304 |
| Melon Piel de Sapo 1 | 0.77 | 0.97 | 0.86 | 246 |
| Tomato 3 | 0.94 | 0.89 | 0.92 | 246 |
| Cherry Rainier 1 | 0.79 | 0.76 | 0.77 | 246 |
| Cherry 2 | 0.95 | 1.00 | 0.98 | 246 |
| Tomato 1 | 0.99 | 0.94 | 0.96 | 246 |
| Strawberry Wedge 1 | 0.80 | 0.70 | 0.75 | 246 |
| Peach 2 | 0.90 | 1.00 | 0.95 | 246 |
| Walnut 1 | 0.69 | 1.00 | 0.82 | 249 |
| Accuracy |  |  | 0.87 | 23619 |
| Macro Avg | 0.88 | 0.87 | 0.87 | 23619 |
| Weighted Avg | 0.88 | 0.87 | 0.87 | 23619 |



## GradientBoostedTrees – Xgboost

Cei mai buni hiper-parametri XGBClassifier: {'n\_estimators': 200, 'max\_depth': 6, 'learning\_rate': 0.1}

Accuracy XGBClassifier: 0.8004149201913714

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Class | Precision | Recall | F1-Score | Support |
| Grape Blue 1 | 0.96 | 0.98 | 0.97 | 328 |
| Plum 3 | 0.80 | 0.94 | 0.86 | 304 |
| Melon Piel de Sapo 1 | 0.76 | 0.86 | 0.81 | 246 |
| Tomato 3 | 0.81 | 0.87 | 0.84 | 246 |
| Cherry Rainier 1 | 0.82 | 0.62 | 0.70 | 246 |
| Cherry 2 | 0.83 | 0.94 | 0.88 | 246 |
| Tomato 1 | 0.93 | 0.90 | 0.91 | 246 |
| Strawberry Wedge 1 | 0.56 | 0.66 | 0.61 | 246 |
| Peach 2 | 0.90 | 1.00 | 0.95 | 246 |
| Walnut 1 | 0.69 | 0.94 | 0.79 | 249 |
| Accuracy |  |  | 0.80 | 23619 |
| Macro Avg | 0.81 | 0.80 | 0.80 | 23619 |
| Weighted Avg | 0.81 | 0.80 | 0.80 | 23619 |

