Etapa 1

Extragerea de atribute (Feature Extraction)

Am implementat un flux de extragere a atributelor folosind două metode principale: descriptori ORB și reducerea dimensionalității cu PCA.

1. Preprocesarea imaginilor

Am redimensionat imaginile la o dimensiune standard de **32x32 pixeli** pentru a uniformiza setul de date, ceea ce reduce complexitatea calculelor și asigură consistență între imagini.

2. Extragerea descriptorilor ORB

Am utilizat ORB pentru a extrage caracteristici semnificative din imagini, cum ar fi colțuri și modele locale.

- **Parametrii aleşi**:
    - `max_keypoints=300:` permite detectarea a până la 300 de puncte de interes pentru fiecare imagine.
    - `edge_threshold=6` pentru controlul sensibilitatii la marginile imaginii
- **Justificare**: ORB este rapid și eficient pentru imagini în tonuri de gri, ceea ce îl face potrivit pentru seturi mari de date. Este rezistent la modificări de scalare sau rotație.

3. Clustering cu KMeans (Bag of Words)

Descriptorii extraşi au fost grupați în **500 clustere** utilizând KMeans, pentru a construi un vocabular vizual.

- Descriptorii din toate imaginile au fost concatenați într-un set comun, iar KMeans a fost aplicat pentru a învăța centrele clusterelor.
- Fiecare imagine a fost reprezentată ca o histogramă bazată pe frecvența clusterelor corespunzătoare descriptorilor săi.
- **Justificare**: Bag of Words oferă o reprezentare compactă și numerică a imaginilor, reducând complexitatea analizei ulterioare.

4. Reducerea dimensionalităţii cu PCA

Am aplicat PCA pentru a reduce dimensiunea histogramelor la **20 de componente principale**.

- Am standardizat datele înainte de PCA pentru a avea o medie de 0 și o varianță unitară.
- **Justificare**: PCA reduce redundanța dintre caracteristici, îmbunătățind eficiența modelelor și explicând ~95% din variabilitatea datelor cu mai puține atribute.

5. Selecţia atributelor

Am utilizat metoda **VarianceThreshold** pentru a elimina caracteristicile cu varianță scăzută (sub 0.01). Aceasta păstrează doar caracteristicile relevante pentru clasificare.

- Selecția atributelor ajută la reducerea zgomotului și a timpului de procesare, îmbunătățind performanța modelului final.

Rezultate și observații

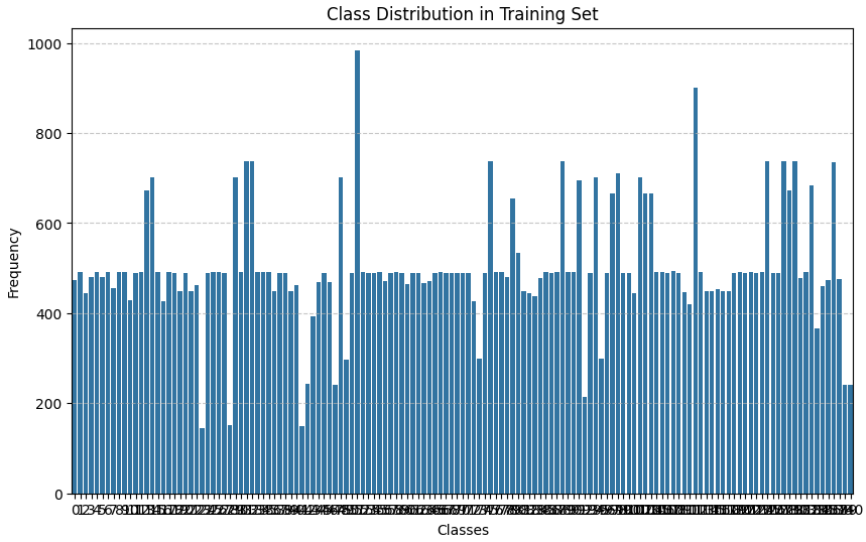
- **Dimensiuni finale**: După fluxul de preprocesare, dimensiunea setului de antrenament a fost redusă semnificativ la *(n_samples, 10)*, unde 10 este numărul de caracteristici finale.
- Metodele alese oferă o bună echilibrare între precizie și eficiență, reprezentând bine imaginile cu un model numeric compact și optimizat.

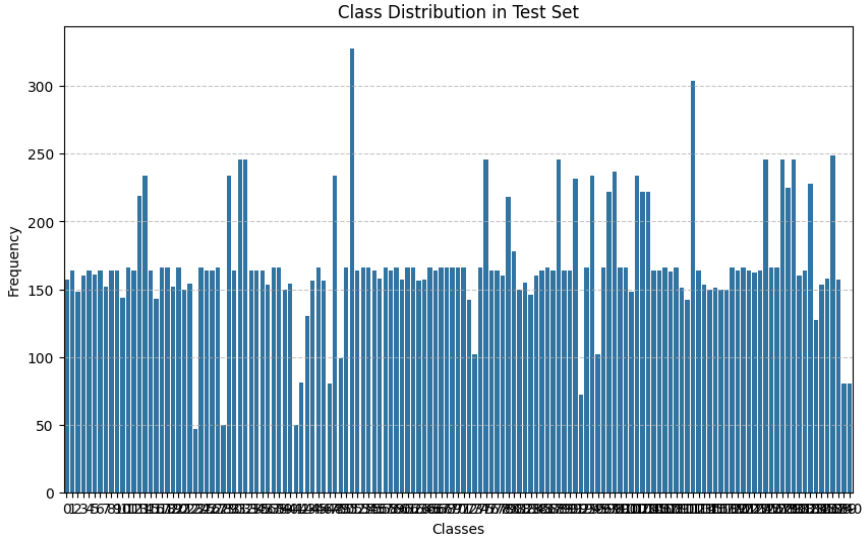
Acest flux poate fi extins sau ajustat în funcție de performanțele modelelor de clasificare ulterioare.

Vizualizarea atributelor extrase

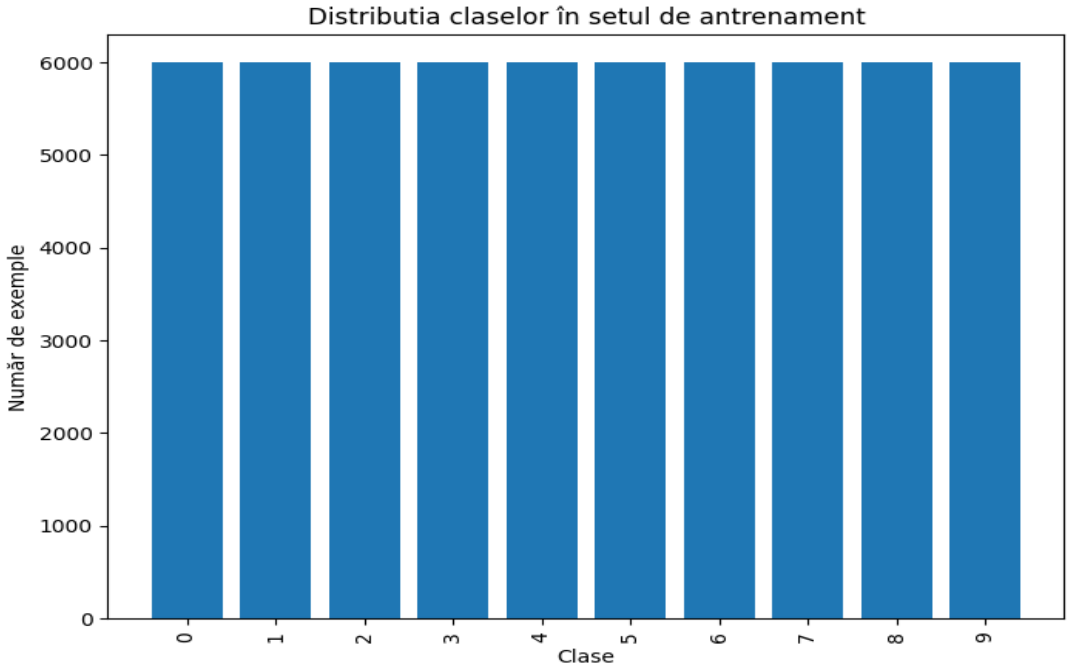
Distributia claselor

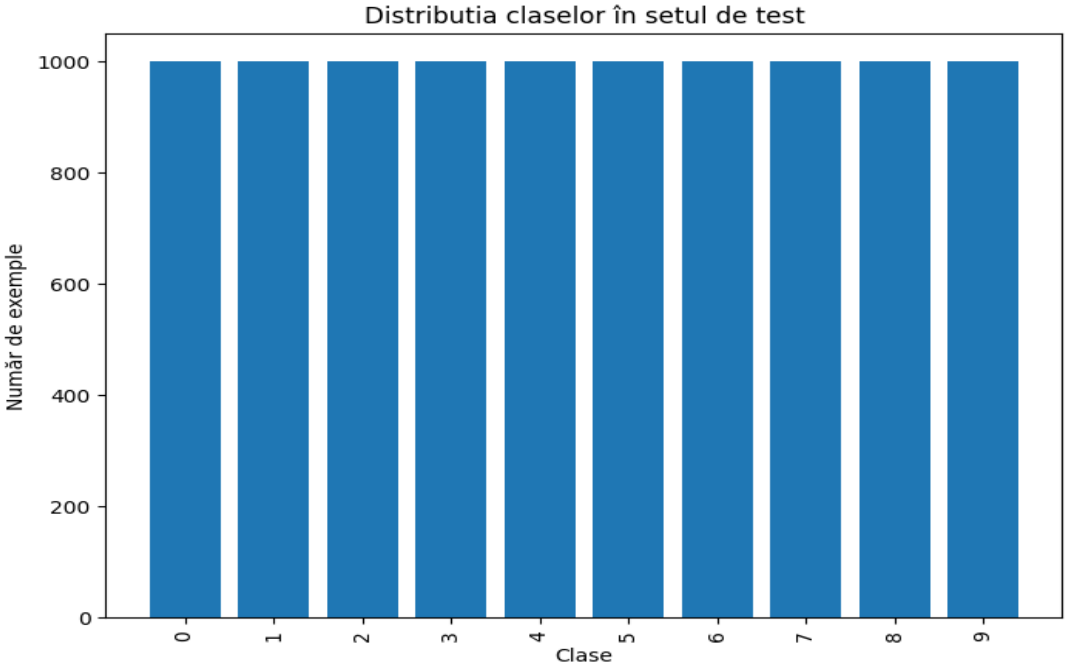
Fruits





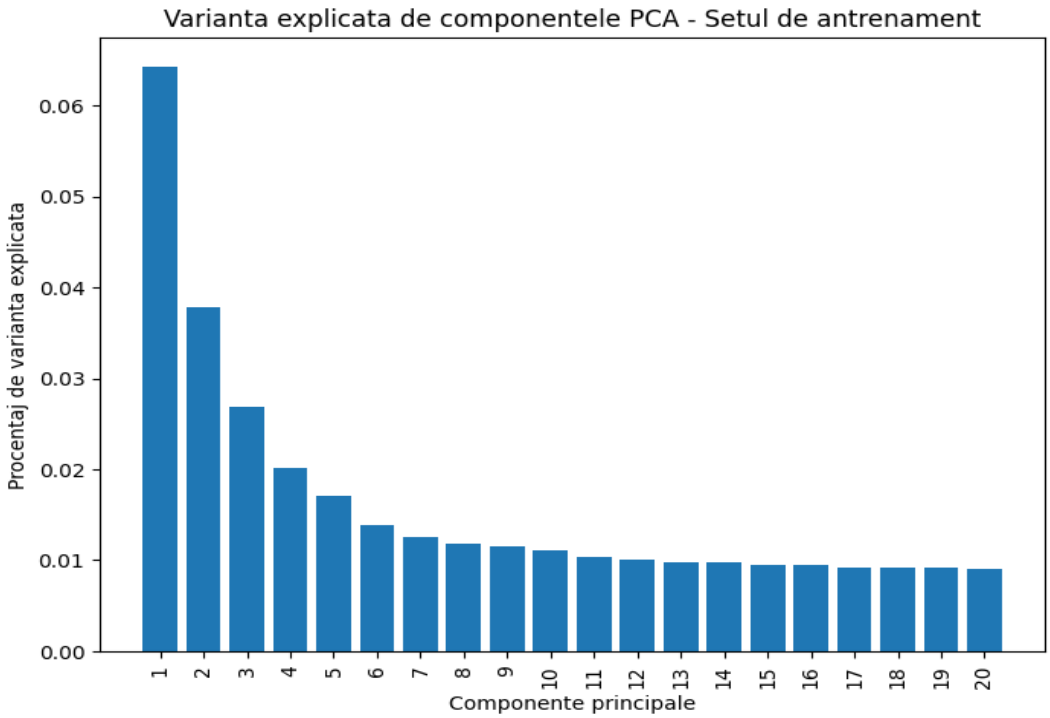
Fashion


Distributia claselor în setul de antrenament
Număr de exemple
0
1000
2000
3000
4000
5000
6000
0
1
2
3
4
5
6
7
8
9
Clase

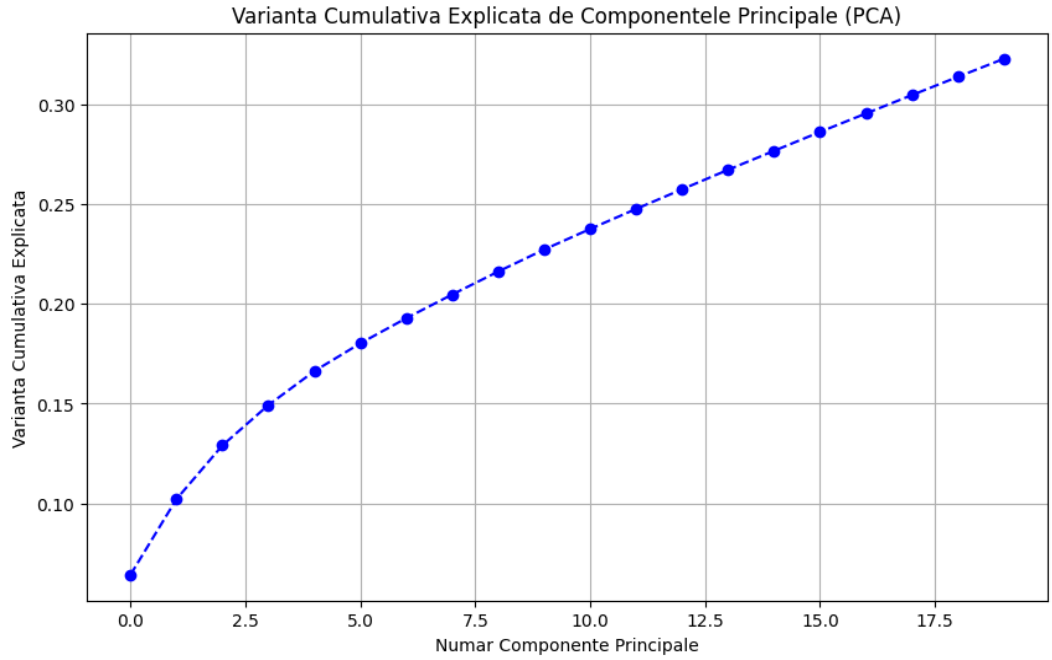

Distributia claselor în setul de test
Număr de exemple
0
200
400
600
800
1000
0
1
2
3
4
5
6
7
8
9
Clase

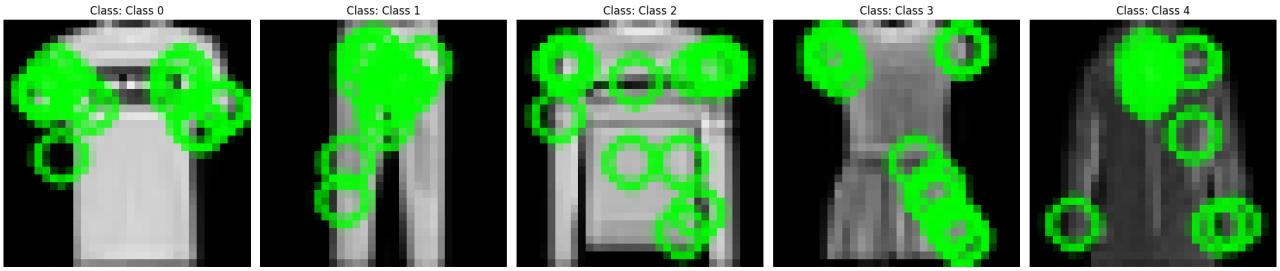
VIZUALIZARE CANTITATIVA SI CALITATIVA

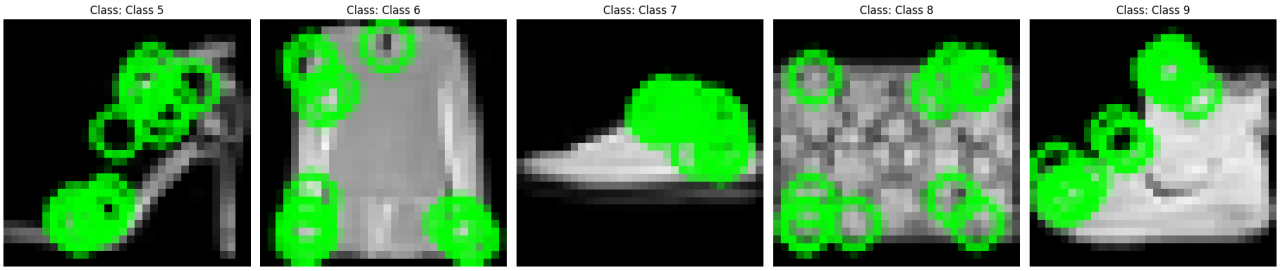
FASHION



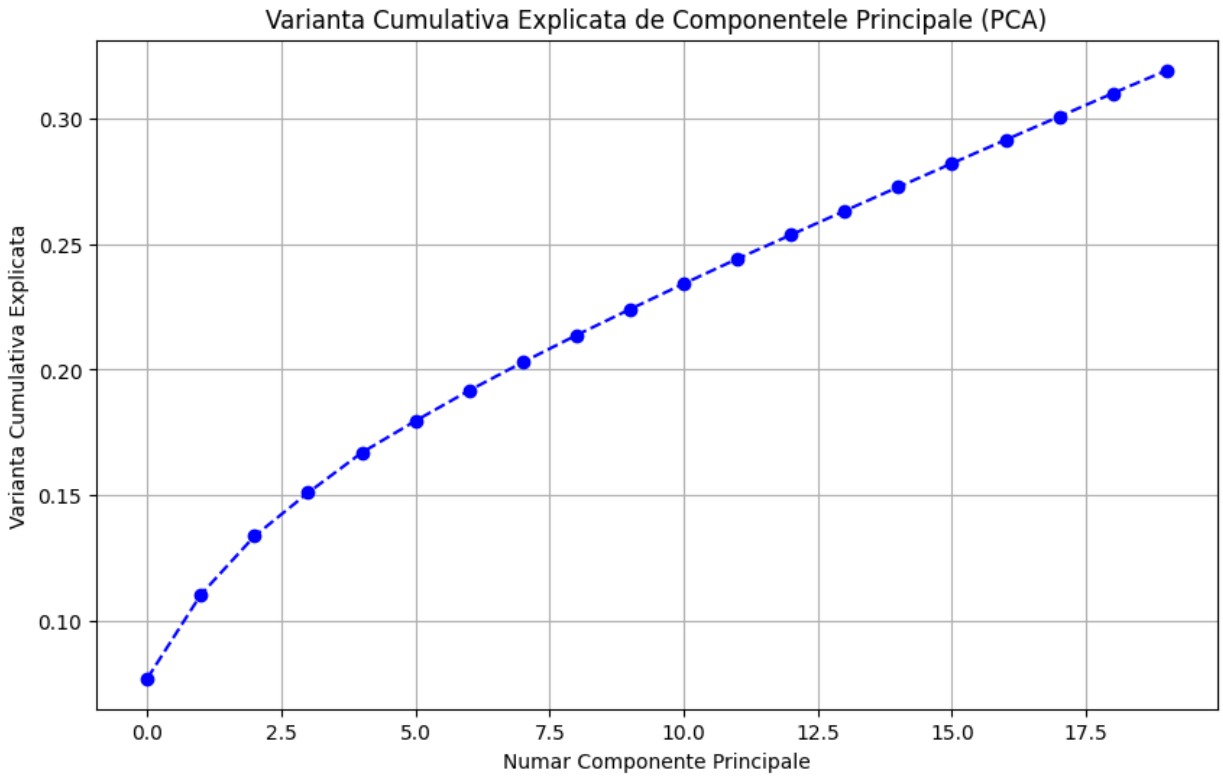
Observăm că prima componentă principală explică o mare parte din varianța datelor, urmată de componentele a doua și a treia care încă contribuie semnificativ, dar ponderea lor scade rapid. Restul componentelor explică doar o fracțiune foarte mică din varianța totală.

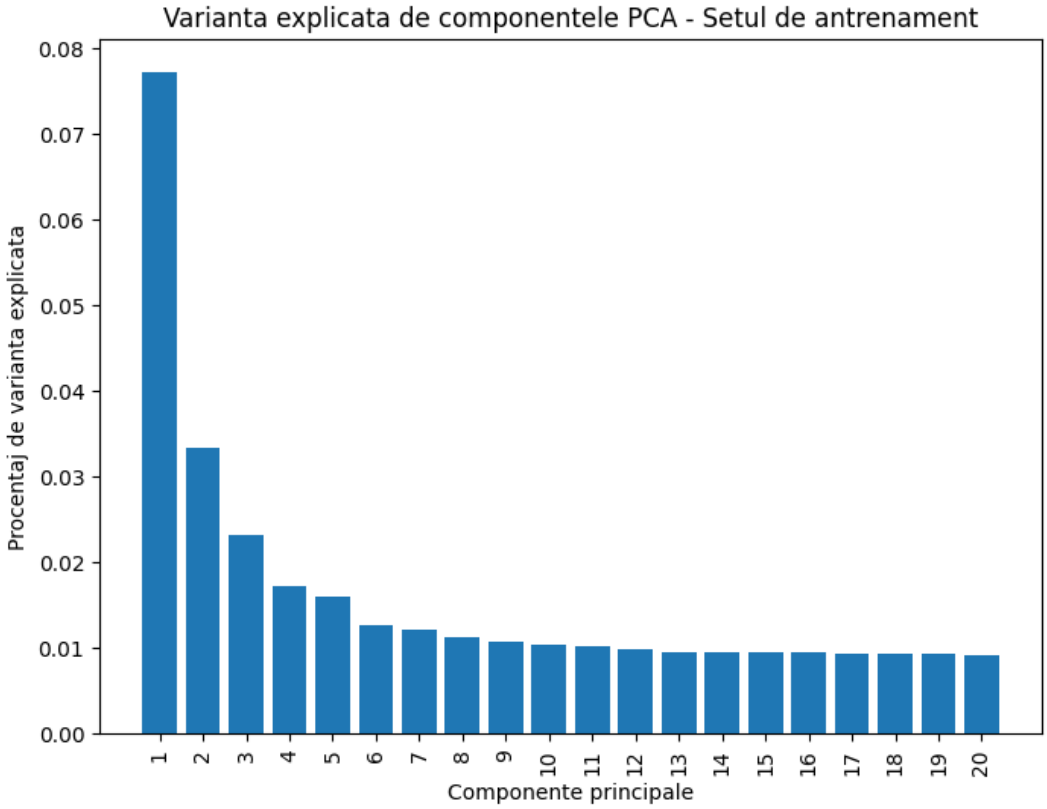


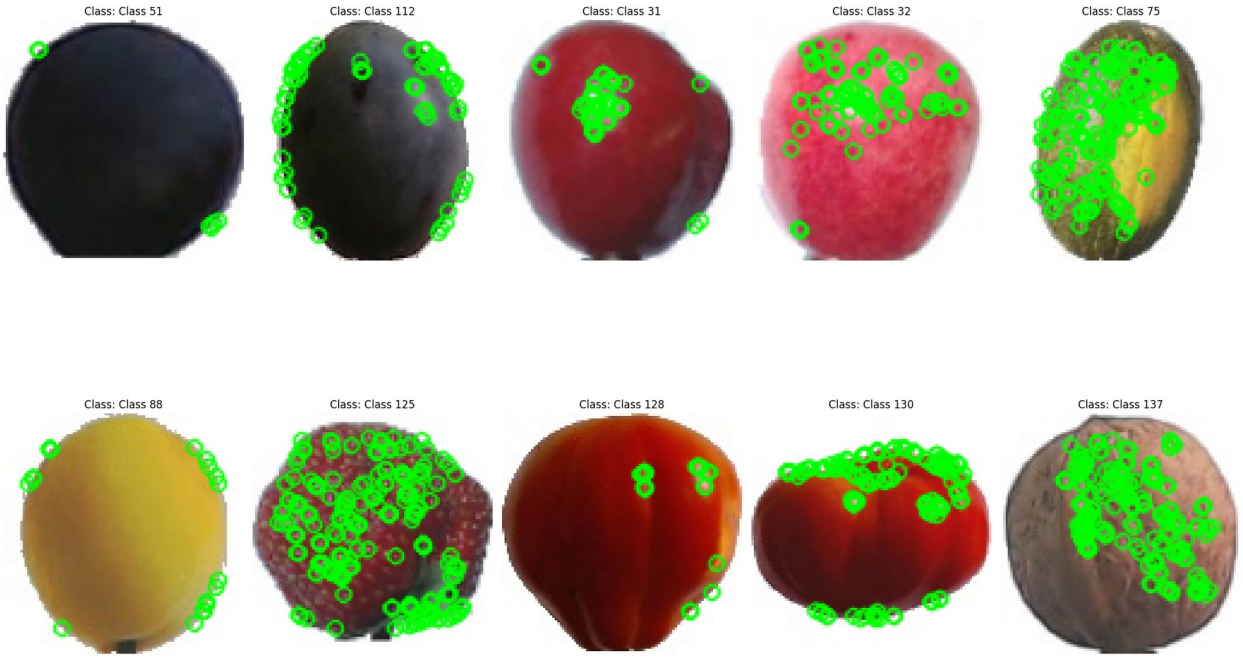

Class: Class 0
Class: Class 1
Class: Class 2
Class: Class 3
Class: Class 4


Class: Class 5
Class: Class 6
Class: Class 7
Class: Class 8
Class: Class 9

FRUITS


Varianta Cumulativa Explicata de Componentele Principale (PCA)
Varianta Cumulativa Explicata
0.10
0.15
0.20
0.25
0.30
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Numar Componente Principale


Varianta explicata de componentele PCA - Setul de antrenament
Procentaj de varianta explicata
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Componente principale


Class: Class 51
Class: Class 112
Class: Class 31
Class: Class 32
Class: Class 75
Class: Class 88
Class: Class 125
Class: Class 128
Class: Class 130
Class: Class 137