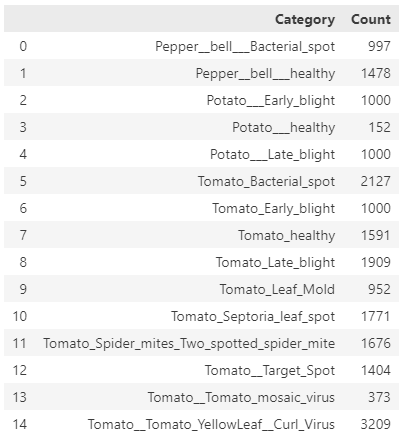
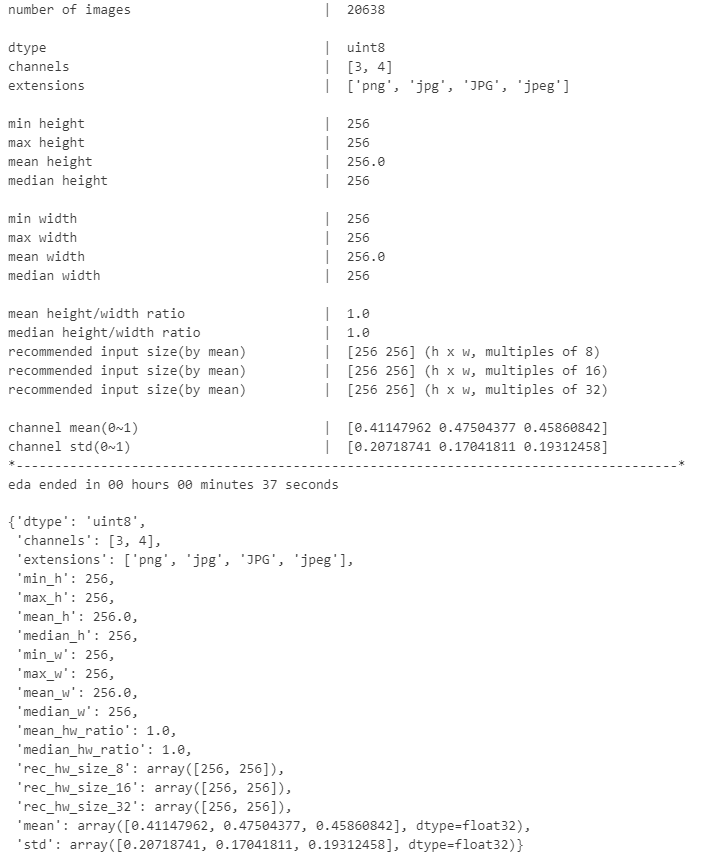
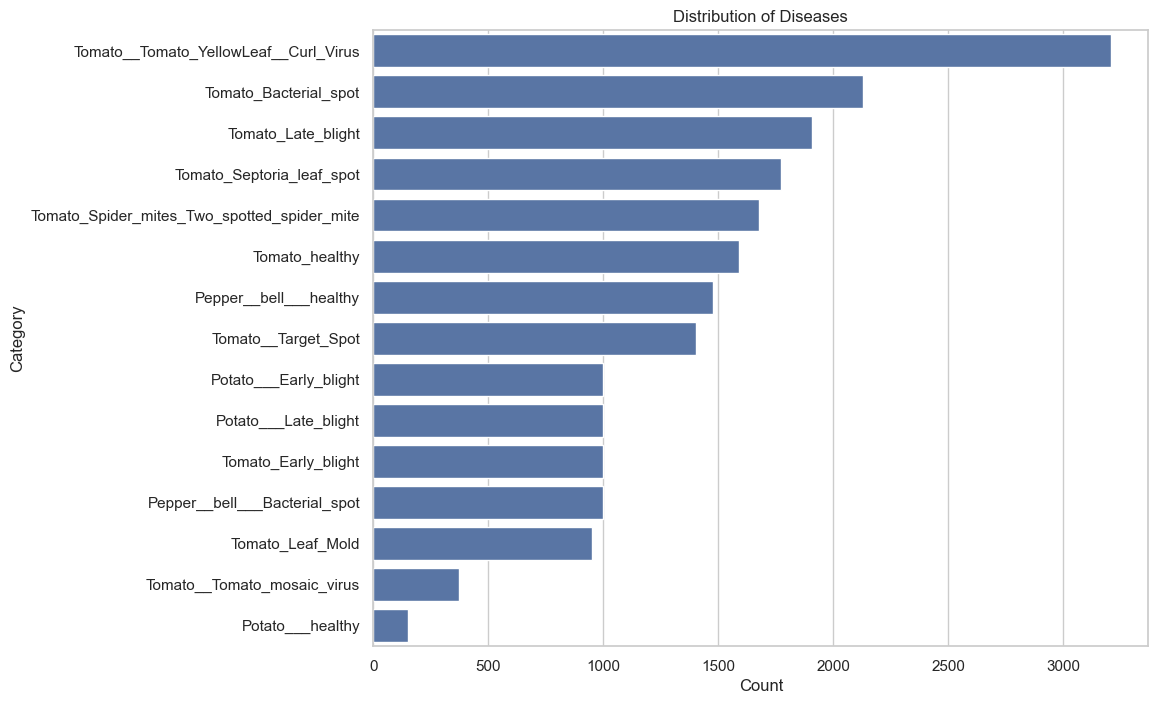
# Laboratorul 2

1. Analiza exploratorie a datelor

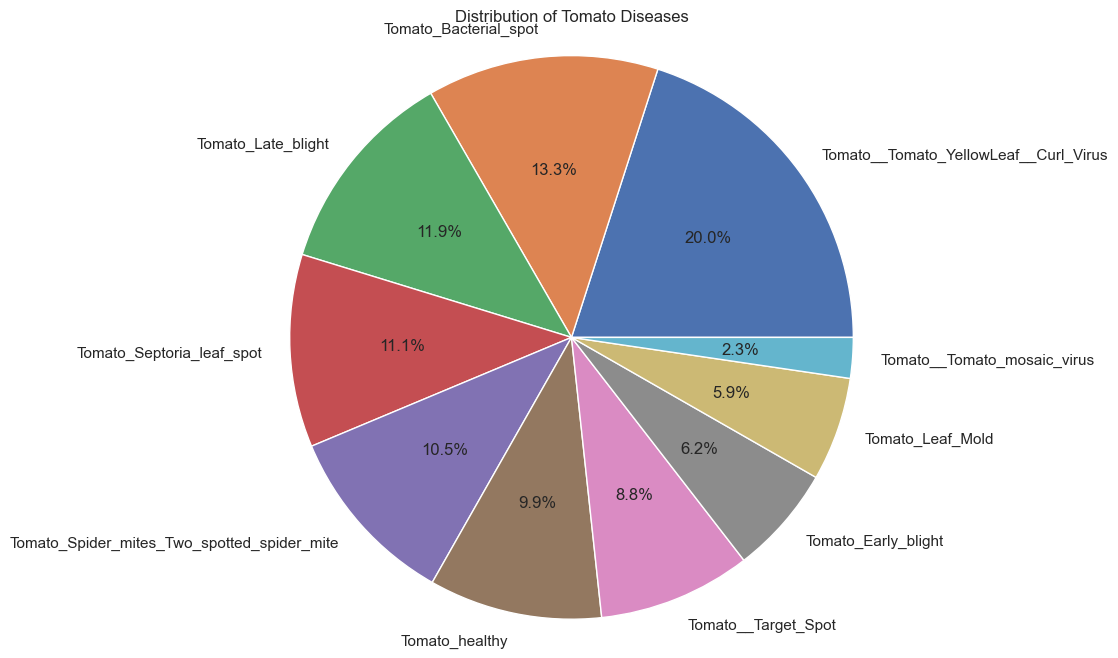
Setul de date folosit are la bază [acest](https://www.kaggle.com/datasets/emmarex/plantdisease/code) set de date, ce conține imagini cu frunze bolnave provenite de la ardei, cartofi și roșii. În urma EDA, am decis ca modelul nostru să fie antrenat pentru detecția frunzelor bolnave pentru roșii, așa că setul nostru de date a fost restrâns la aceste imagini.

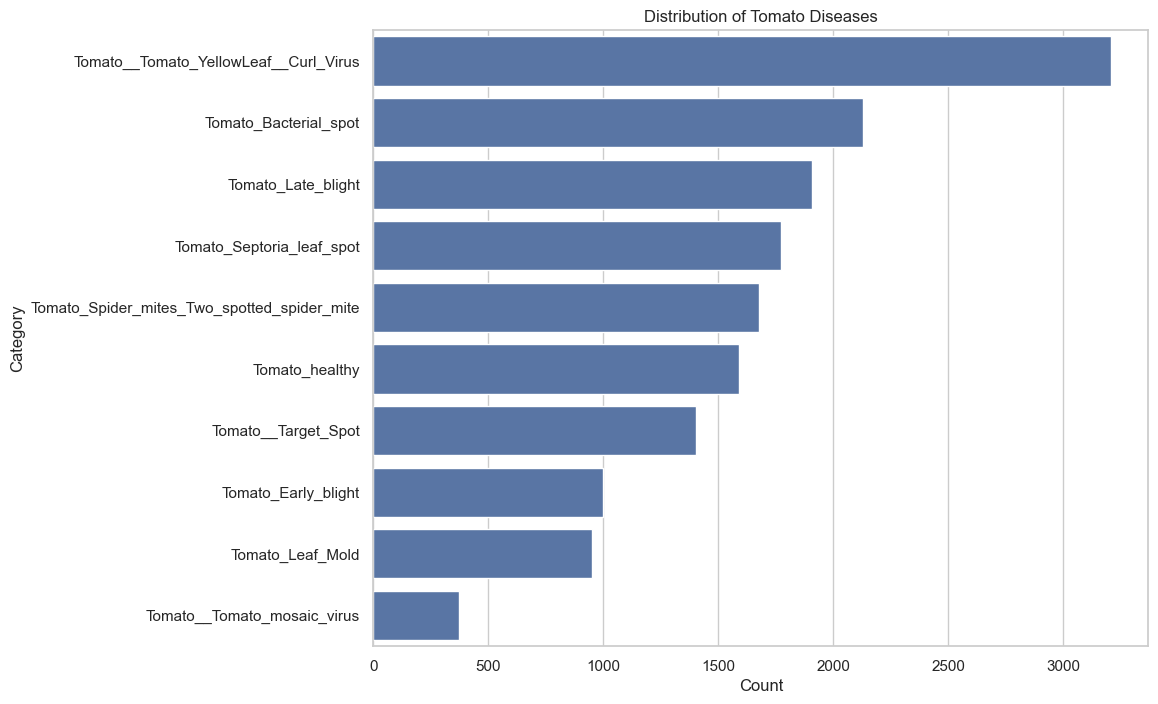
Setul de date inițial conținea 20638 de imagini, toate având dimensiunea de 256x256 de pixeli. Existau diferențe considerabile între numărul de imagini cu roșii pe care le conținea setul de date și numărul de imagini cu ardei sau cartofi, așa că am decis să ne antrenăm modelul doar pentru a detecta frunze de roșii bolnave.



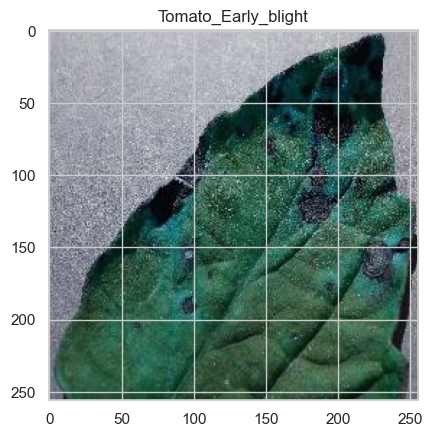
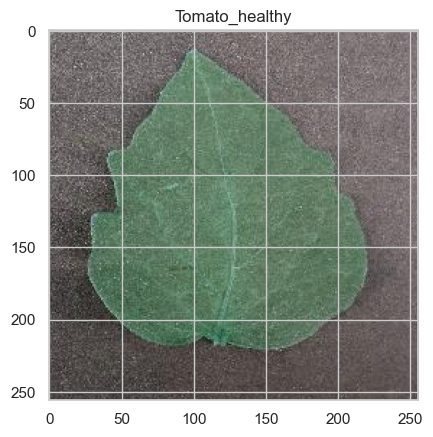


După ce am restrâns setul de date, am observat în continuare faptul că datele trebuiau augmentate, astfel încât modelul să aibă acces la același număr de date din fiecare categorie pentru a învăța cât mai bine să detecteze corect bolile. De asemenea, am vrut să vedem care este proporția de frunze sănătoase/bolnave, iar apoi am decis că frunzele sănătoase vor fi tratate ca o clasă individuală.





Ulterior, am vrut să vizualizăm datele, astfel că am făcut plot la câte un exemplar din fiecare categorie de frunze.



1. Preprocesarea datelor

Primul pas al preprocesării a fost să reducem dimensiunea imaginilor de la 256x256 la 128x128 de pixeli, astfel încât procesarea acestora să fie mai ușoară.

Întrucât în timpul EDA, am observat că setul de date este imbalansat, am decis să augmentăm datele, până când obținem un număr egal de imagini (3209, clasa maxima nu are augmentare) pentru fiecare categorie de frunze din setul de date. Augmentarea a fost realizată alegând o imagine random dintr-o categorie și alegând random daca ii aplicăm un flip (vertical/orizontal) și rotind imaginea cu un unghi random între -30 și 30 de grade.