314CA Biciusca Tudor-Andrei

Galaxy Classification Project

This project classifies galaxy images into four types: Elliptical, Spiral, Barred Spiral, and Other/Unclassified.
The classifications are based on a simplified Hubble Sequence.

Overview

Starting from:

1. One folder containing ~16k images of galaxies in 424x424 resolution, each with it's own asset_id
2. A csv file that links asset_id to obj_id which is the standard way to catalog galaxies
3. A csv file containing obj_id and a lot of features including the classification of that galaxy

The process:

1. Link the asset_id to the classification
2. The dataset uses a lot of classifications (very specific ones) so for the purposes of this project we group them in the 4 big ones mentioned above
3. Crop images to keep only the main galaxy (otherwise the photos are very noisy and cluttered with other galaxies/stars/dust)
4. Equalize the dataset by applying effects (rotations, shears, mirroring...)
5. Extracts numerical features from each cropped image using both math based features and a CNN.
6. Select the most important features (between 15 and 20 give the best results as far as I've tested)
7. Trains a Random Forest classifier on those features.

Detailed Explanation

Data Preparation

* Original images: `dataset/raw/images/`
* CSV mapping: `dataset/processed/augmented.csv` with columns `image_path,galaxy_class`

Augmentation

I wrote a script to balance classes by flipping, rotating, and adjusting brightness.

Cropping

I use OpenCV to:

1. Blur and threshold the image.
2. Find contours and pick the one nearest the center.
3. Crop around that contour and save to `dataset/processed/cropped/`.

Feature Extraction

I extract:

* 8 classical features (concentration, ellipticity, gini, M20, asymmetry, smoothness, mean intensity, edge density)
* 10 HOG features
* 6 RGB color histogram bins
* 50 CNN embedding features (ResNet50)

Results are saved to `dataset/processed/all_features_cropped.csv`.

Feature Selection and Training

1. Load features and split into train/test (80/20).
2. Train a Random Forest to get feature importances.
3. Pick the top 15 features and retrain a second Random Forest.
4. Evaluate on the test set and print metrics.

Results

* Final accuracy: ~65%
* Classes still confused: Spirals vs Barred Spirals.

Problems and Fixes

* Too many types of galaxy some of them having only 1 (one) example : Grouping galaxies based on their prefix
* Predictions preferred Elliptical Galaxies due to the big number of data on them : Added image effects to equalize the number of photos.
* Very low original accuracy : Better feature selection using a CNN and more features in general (went from 8 to 15)
* Slow processing: added tqdm progress bars to make sure we don't find out the program is stuck after 1 hour (learned from experience).
* Noisy images: cropped central galaxy before feature extraction.

Next Steps

* Improve image preprocessing (deblurring, images are still not cropped optimally, galaxies that overlap are not separated properly).
* Try better features to distinguish between Barred Spiral and Spiral (LBP, Fourier features).
* Test other models (XGBoost, MLP).

Other Details

* Time to extract the features ~ 1:30 (i have no gpu and bad cpu on the machine i worked on)
* The modest results are part due to my inexperience but also because of the low quality data we are working with

```
Classification Report: Top 15 Features
                 precision     recall  f1-score   support

        Barred Spiral        0.60    0.60    0.60    1165
        Elliptical      0.65    0.72    0.68    1165
Other/Unclassified        0.81    0.89    0.85    1164
        Spiral          0.53    0.42    0.47    1165

        accuracy                         0.66    4659
        macro avg       0.65    0.66    0.65    4659
        weighted avg         0.65    0.66    0.65    4659
```

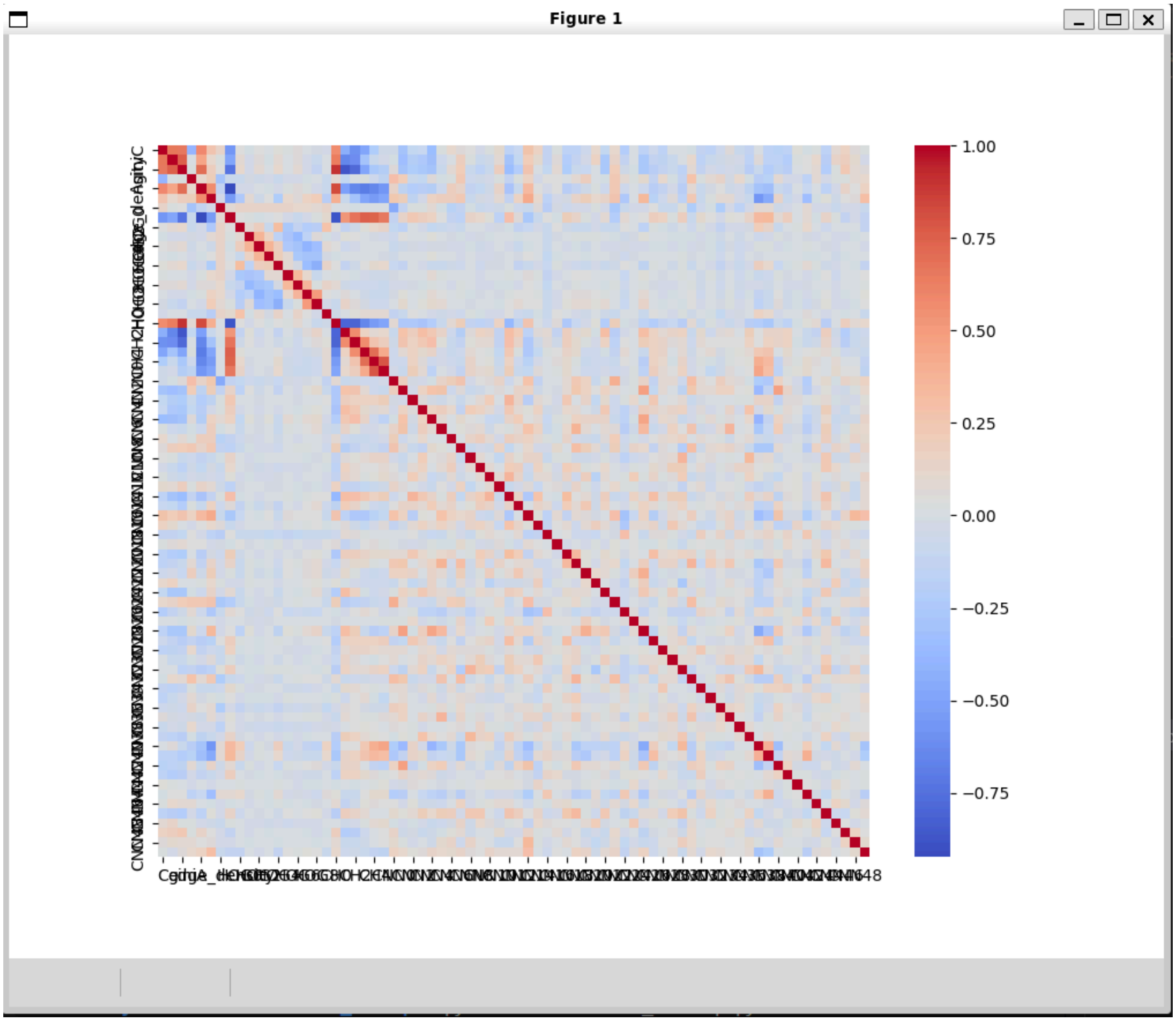
Classification Report: Full features

```
                 precision      recall  f1-score   support

        Barred Spiral           0.57    0.60    0.59    1165
        Elliptical      0.64    0.73    0.68    1165
Other/Unclassified              0.79    0.89    0.84    1164
        Spiral          0.54    0.37    0.44    1165

        accuracy                        0.65    4659
        macro avg       0.64    0.65    0.64    4659
        weighted avg            0.64    0.65    0.64    4659
```

As we can see the results are slightly better when we only use part of the features.



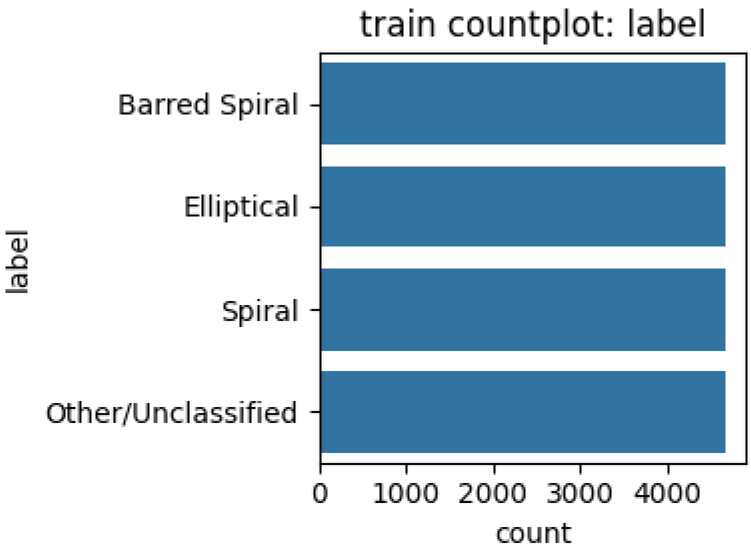
Testing for the optimal number of features to be chosen

```
          accuracy  macro_f1
K
5   0.578880  0.571518
10  0.633827  0.626221
15  0.656579  0.648714
20  0.658725  0.650257
30  0.653788  0.644841
50  0.648208  0.638550
```

From these results we conclude that the optimal number of features is ~20.

We don’t have any missing data values.

```
--- EDA for test ---
            missing_count  missing_pct
label                   0          0.0
C                       0          0.0
ellipticity             0          0.0
gini                    0          0.0
M20                     0          0.0
...                   ...          ...
CNN45                   0          0.0
CNN46                   0          0.0
CNN47                   0          0.0
CNN48                   0          0.0
CNN49                   0          0.0
```

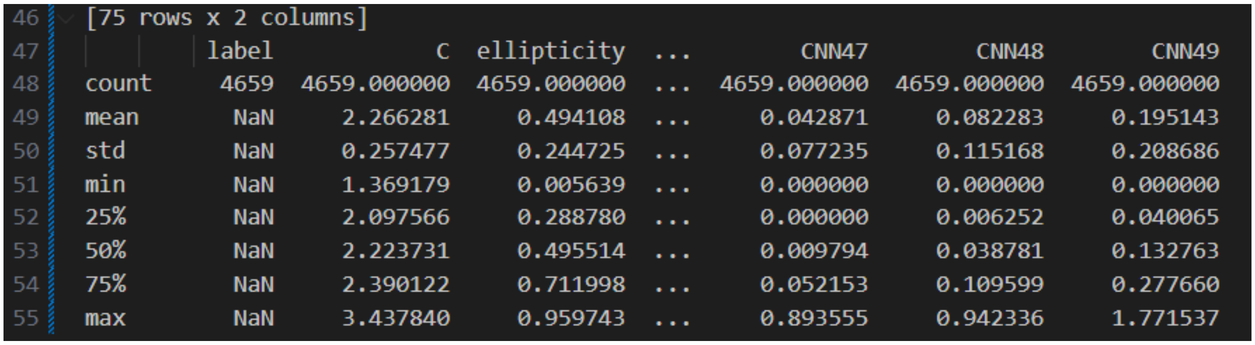


We can see that our number of images is almost equal, the difference is at most 5 and that happened in the cropping part of preprocessing. Images that had a very small galaxy were deleted but I made the threshold small enough so our data won’t be skewed.

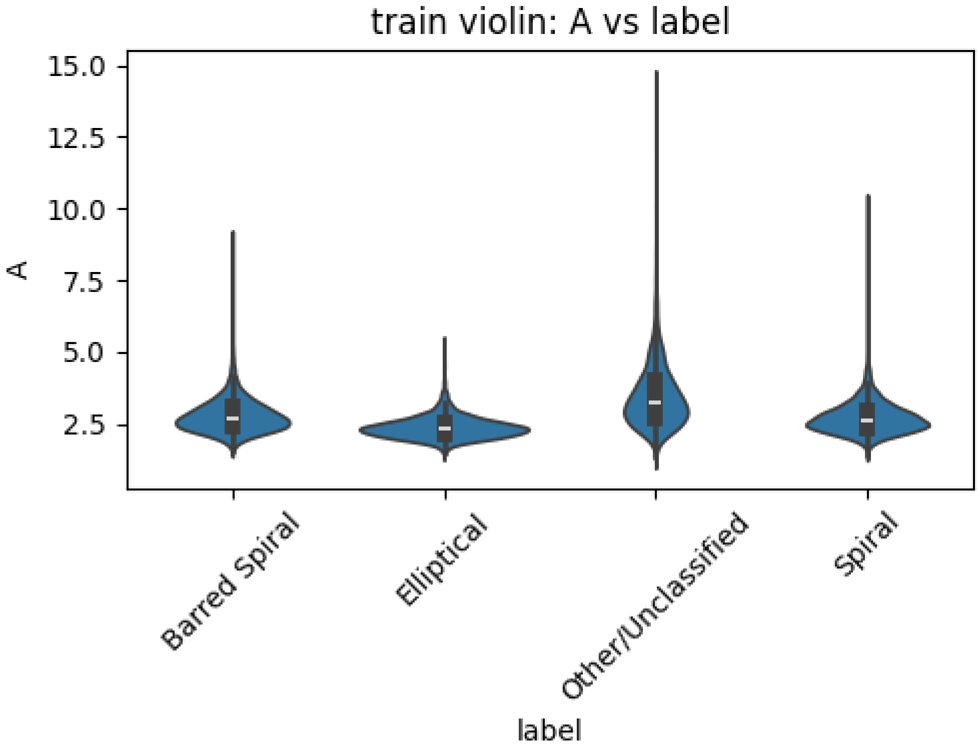
Detailed count:

```
label,count
Barred Spiral,4661
Elliptical,4660
Spiral,4659
```

Other/Unclassified,4656

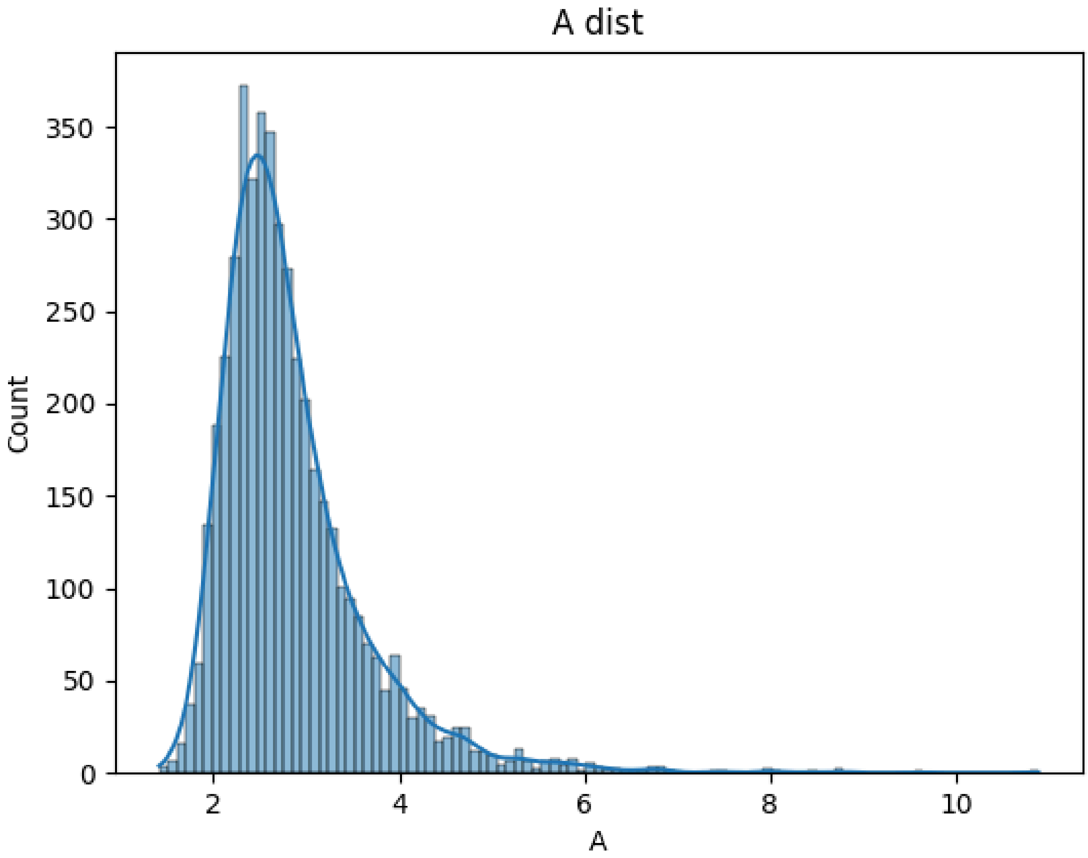


```
[75 rows x 2 columns]
      label            C  ellipticity  ...        CNN47        CNN48        CNN49
count  4659  4659.000000  4659.000000  ...  4659.000000  4659.000000  4659.000000
mean    NaN     2.266281     0.494108  ...     0.042871     0.082283     0.195143
std     NaN     0.257477     0.244725  ...     0.077235     0.115168     0.208686
min     NaN     1.369179     0.005639  ...     0.000000     0.000000     0.000000
25%     NaN     2.097566     0.288780  ...     0.000000     0.006252     0.040065
50%     NaN     2.223731     0.495514  ...     0.009794     0.038781     0.132763
75%     NaN     2.390122     0.711998  ...     0.052153     0.109599     0.277660
max     NaN     3.437840     0.959743  ...     0.893555     0.942336     1.771537
```



Observations :

Similarities: "Barred Spiral," "Elliptical," and "Spiral" have overlapping distributions (mostly 5.0 to 7.5), meaning "Asymmetry" alone might not easily separate these types.
Differences: "Other/Unclassified" stands out with a higher peak (7.5 to 10.0) and some unique high values (up to 15.0). This suggests "Asymmetry" could be especially helpful for identifying this group.


A dist
Count
A
0
50
100
150
200
250
300
350
2
4
6
8
10