Prepared by group 10

Project Demo

F20 DL

Adam Aboushady, Sri Sai Vaishnavi Chintha, Mustansir Eranpurwala, Ihsan Fazal, Janya Rathnakumar

Project Topic

Galaxy Morphology and Quenching State Classification using ML models

Problem Context:

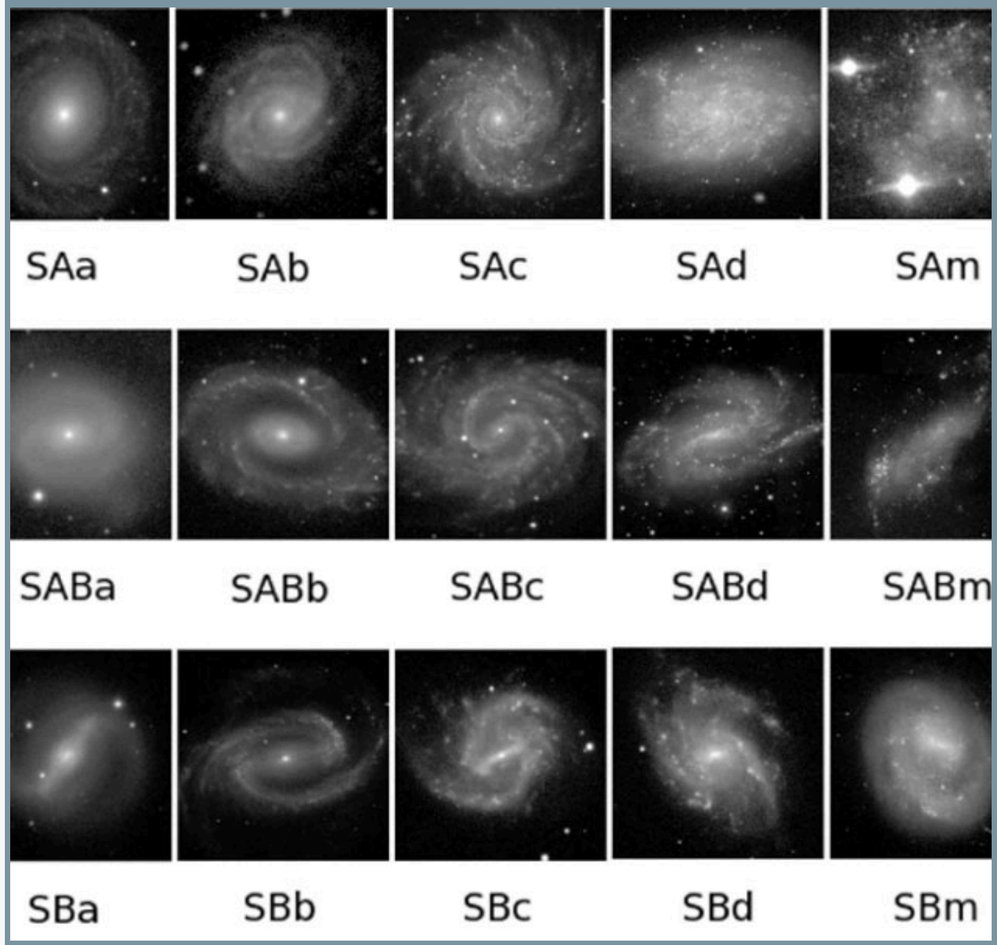
- Understanding galaxy evolution requires accurate classification of morphology (shape) and quenching state (star formation activity).
- Traditional manual labeling (e.g., Galaxy Zoo) is not scalable due to the growth of data from surveys like SDSS.

Research Questions:

- How do different data representations and ML approaches compare in effectiveness for galaxy morphology classification?
- Can the quenching state be reliably classified using image-based and photometric feature-based deep learning approaches?

Model Selection Strategy:

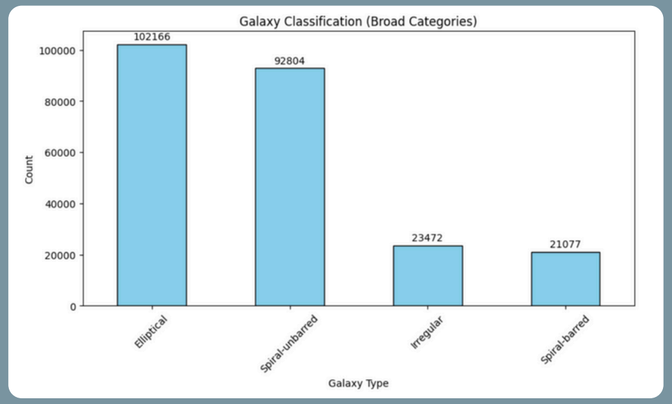
- Classical ML (Decision Trees, KNN, Naive Bayes) → Morphology classification; Galaxy Zoo features derived from human consensus votes, capture structured, discriminative information about galaxy shape.
- MLP (Photometric features) → Morphology and quenching classification; photometric numerical values capture galaxy color, brightness, and spectral properties linked to star formation and morphology.
- CNN (Image data) → Morphology and quenching classification; learns spatial patterns (e.g., spiral arms, bars) and visual cues related to evolutionary stages from raw pixels.

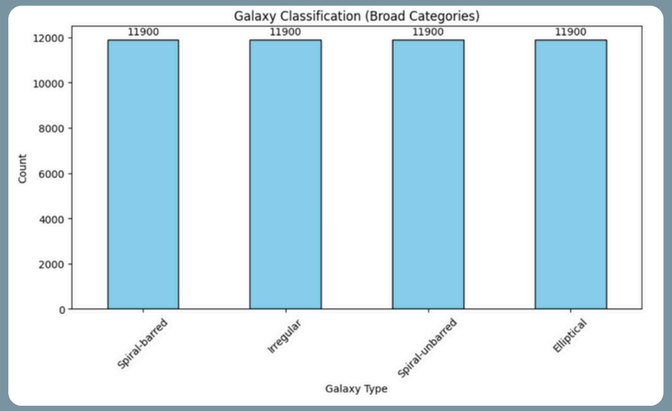


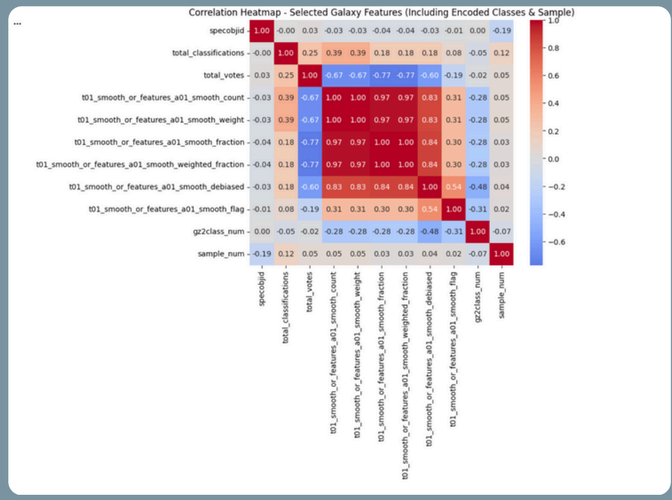
Data Description & Pre-processing

Galaxy Zoo 2 (Tabular, SDSS DR7/DR8):

- Dataset size: 243,500 galaxies.
- The dataset contains 233 features: IDs, coordinates, total classifications and votes, per-task vote counts, weighted votes, and gz2class, which is a shorthand string representing the morphology for each galaxy(Sa, Sb, Ei, SBb(r)).
- Added a new feature broad_class, which assigns each galaxy to one of four morphological classes based on gz2class: Elliptical, Spiral-unbarred, Spiral-barred, and Irregular.
- Elliptical and Spiral-unbarred galaxies dominated the dataset, while Spiral-barred and Irregular were much smaller (fig 1). So we balanced by downsizing.
- After balancing the tabular dataset, we matched it with the SDSS image dataset using the specobjid feature (DR8 spectrum object ID). Out of ~400k images, 47k galaxies had a corresponding entry in our balanced tabular subset, which we used as the final image dataset for experiments.
- Galaxy Zoo 2 Tabular Link: Galaxy Zoo Data Release
- Key Reference: Galaxy Zoo 2: Detailed morphological classifications for 304,122 galaxies (Willett et al., 2013, DOI:10.1093/mnras/stt1458)







Preprocessing:

Google Colab
google.com

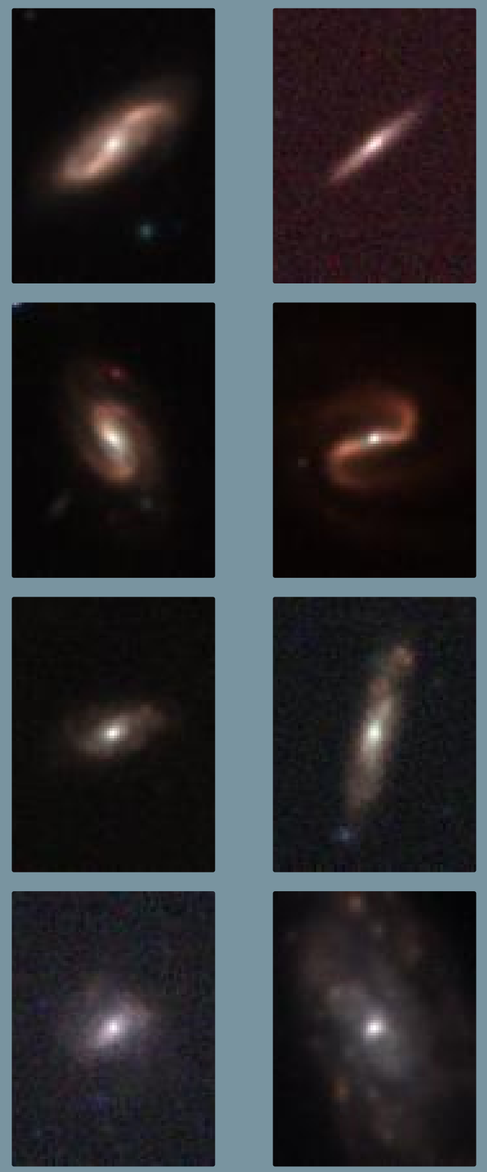
Data Description & Pre-processing

SDSS Image Dataset (NERSC SSL release):

- Dataset size: 399,982 images
- Each galaxy in the dataset is provided as an image in the five standard SDSS photometric bands (ugriz). These bands represent different wavelength ranges of light captured from the galaxy : **u-band** (354 nm - ultraviolet light), **g-band** (477 nm - green/blue visible light), **r-band** (623 nm - red visible light), **i-band** (763 nm - near-infrared), **z-band** (913 nm - deep near-infrared)
- For every galaxy, the dataset gives us five aligned images (one per band), capturing how the galaxy looks across different parts of the spectrum. All the images are of the size 107x107 pixels.
- Attributes “death_status” and “broad_class” were introduced into the dataset to enable classification. The “death_status” attribute was classified into 3 categories - Dead, Alive and Transitional to indicate the stage of life of the galaxy. These categories were derived from these photo metric values.
- The image dataset was matched with the the tabular dataset after normalisation and removal of outliers. The “specObjID” attribute was used to match the datasets. The final number of records in the image dataset matched the tabular dataset at 47K.
- SDSS Images source link: NERSC SDSS Dataset

Preprocessing:

Google Colab
google.com



ML Models – Results & Insights

To perform an initial assessment of morphology classification using the Galaxy Zoo 2 tabular dataset, we implemented three ML models, Naive Bayes, k-NN, and Decision Trees. These models allowed for the evaluation of simple feature driven approaches before introducing complex neural networks.

Model	Train Accuracy	Test Accuracy	Key Insight
Decision Tree	98.87%	98.60%	Best performer, interpretable, learns strong feature rules.
Naive Bayes	83.76%	83.52%	Fast baseline, limited by independence assumption.
k-NN	85.76%	84.43%	Stable, simple distance-based classifier.

Design Choices:

- Use of balanced tabular features
- Hyperparameter tuning:
- Naive Bayes: variance smoothing (10^{-12} to 10^{-4})
- k-NN: neighborhood size (k = 5 to 25)
- Decision Tree: depth (5–30), impurity (gini/entropy)

Model Limitations:

- Decision Tree: Sensitive to overlapping or noisy feature regions.
- Naive Bayes: Assumption of feature independence limits accuracy on correlated galaxy features.
- k-NN: Sensitive to feature scaling and struggles when classes overlap.

Neural Networks – Results & Insights

Model	Task	Train Accuracy	Test Accuracy	Key Insight
MLP	Quenching	98.90%	99.34%	Exceptional, reliable, photometric features
CNN	Quenching	79.18%	77.96%	Moderate, visual features insufficient
MLP	Morphology	49.13%	47.27%	Limited, underperformed tabular, simple architecture
CNN	Morphology	66%	64.67%	Failure, lacks spatial information

Design Choices:

- Data inputs: CNN used 107 x 107 x 5 raw images, MLP used derived numeric photometric features
- Hyperparameter tuning:
- CNN: 2 Conv2D layers and 2 dense layers
- MLP: varied network depths (128-64 to 512-256-128), LR = (1e-3)

Model Limitations:

- Data representation: data high dependent on input, complex features needed for structure
- CNN: simple architecture limits raw feature extraction.
- MLP: cannot model spatial/structural complexity from numeric input

Conclusion

Key Findings:

Task	Best Data Type	Best Model	Key Insight
Morphology	Tabular (vote fractions)	Decision Tree	The morphology task was best captured using structured, feature-driven tabular data because human-derived Galaxy Zoo features explicitly encode the key spatial structures of galaxies, such as bars, spiral arms, and ellipticity, which CNNs struggle to extract reliably from raw pixel images without extensive data and complex architectures
Quenching	Photometric (u, g, r, i, z bands)	MLP	Tabular photometric features distill the color, magnitude, and luminosity signals (the true drivers of quenching) into highly efficient numerical indicators, which the MLP models effectively. Raw image CNNs perform worse because irrelevant spatial noise and structural variations interfere with the extraction of the crucial color/magnitude signal

Limitations:

- The CNN's performance was significantly limited by its shallow architecture and high computational cost, preventing it from effectively extracting complex morphological cues like bars and spiral arms from raw image data. For the quenching task, it failed to isolate the crucial color signal from spatial noise, leading to lower accuracy than the feature-based MLP

Contribution:

- The project followed a specialized workflow: Vaishnavi managed tabular data preprocessing and K-NN, Janya implemented the Decision Tree models, Mustansir handled CNN development and image data preprocessing, Adam worked on the MLP, Ihsan implemented Naive Bayes, and the team collaborated on the final documentation and presentation slides.

Thank you