Prepared by group 10

Project Pitch

F20 DL

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

Project Topic

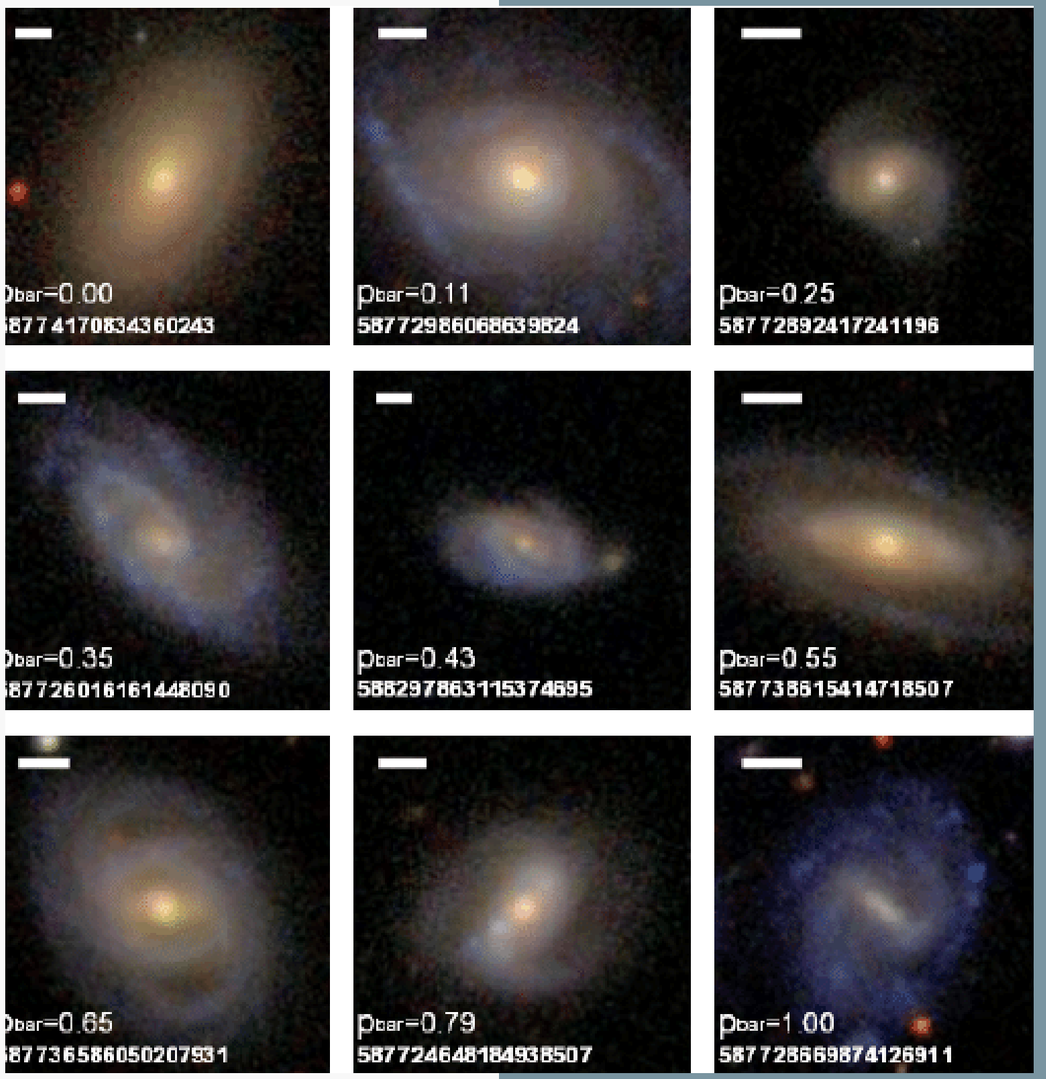
Galaxy Morphology Classification and Quenching State Prediction using ML models

Overview:

- Focus on two astrophysical problems:
  - Galaxy morphological classification (spiral vs elliptical, etc.)
  - Galaxy quenching prediction (whether a galaxy has stopped forming stars)
- Assess and compare predictive performance of:
  - Image-based models
  - Tabular-data models
  - Hybrid approaches (image + tabular)

Why We Picked This Topic:

- Scientific significance – key to understanding galaxy evolution.
- Rich datasets – SDSS & Galaxy Zoo are large, publicly available, and widely used in astronomy + ML research.
- Comparative learning – allows evaluation of image-driven vs tabular-driven ML.



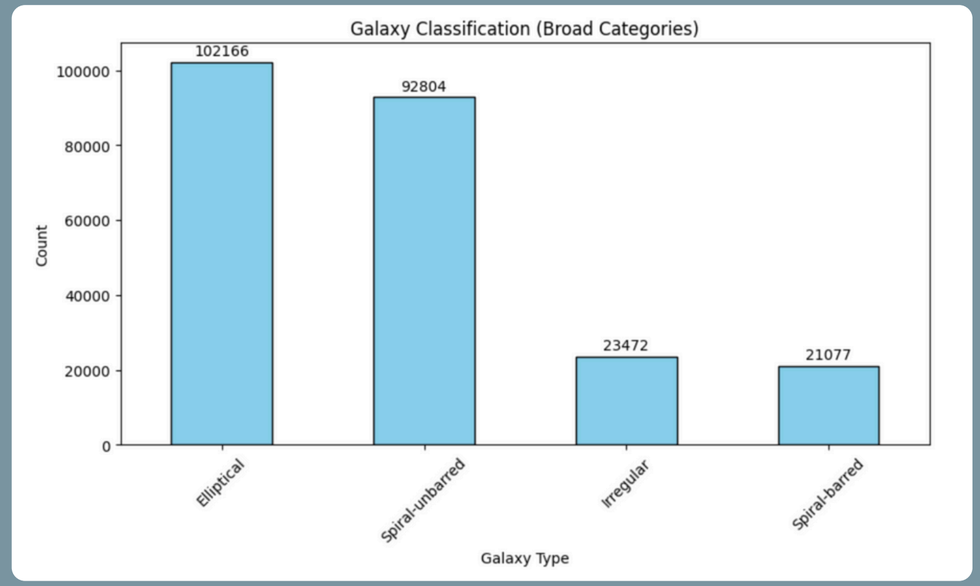
Data Description

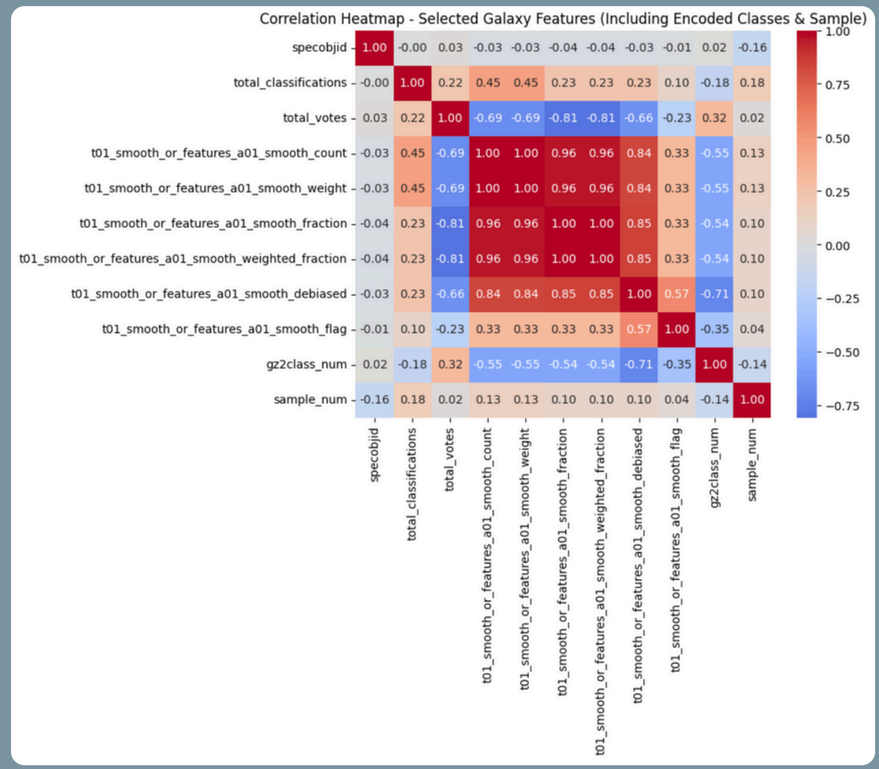
Galaxy Zoo 2 (Tabular, SDSS DR7/DR8):

- Dataset size: 243,500 galaxies.
- Feature scope: The dataset contains 233 features including IDs and coordinates (Nominal/Interval), total classifications and votes (Interval), per-task vote counts, weighted votes, fractions, weighted fractions, debiased fractions (Interval), and binary flags and gz2class, which is a shorthand string representing the most common consensus morphology for each galaxy (Nominal).
- 4 Morphological Classes created on gz2class: Elliptical, Spiral-unbarred, Spiral-barred, and Irregular
- 3 Quenching States Classes created on Morphological classes: Alive (star-forming), Dead (quenched), Intermediate (irregular star formation)
- 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
co google.com





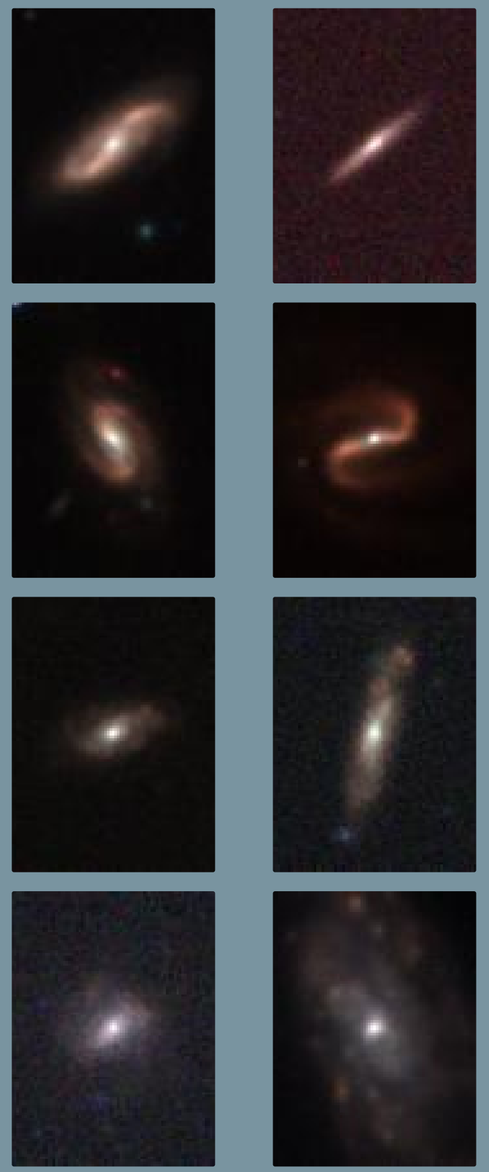
Data Description

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.
- SDSS Images source link: NERSC SDSS Dataset



Preprocessing:

Data Source & Past Applications

Previous ML Applications:

- Hayat et al. (2020): Self-Supervised Representation Learning for Astronomical Images. DOI 10.3847/2041-8213/abf2c7
  - Aim: To show that Self Supervised Learning can pretrain useful galaxy features from raw pixels.
  - Dataset: SDSS Data Release 12
  - Difference: This study only uses image data, tabular data like GZ2 fractions are not compared.
- Dieleman et al. (2015):Rotation-invariant CNNs for galaxy morphology. DOI 10.1093/mnras/stv632
  - Aim: To achieve high accuracy supervised classification of morphology from GZ2 labels.
  - Dataset: Galaxy Zoo 2 images.
  - Difference: Our work will be integrating both image and tabular features, connecting morphology to quenching.
- Domínguez Sánchez et al. (2018): Improving galaxy morphologies for SDSS with Deep Learning. DOI 10.1093/mnras/sty338
  - Aims: Create a sizable, uniform morphology catalog for SDSS
  - Dataset: SDSS Data Release 7
  - Difference: This study only uses image data, tabular data like GZ2 fractions are not compared.
- Smethurst et al. (2017): Galaxy Zoo: the interplay of quenching mechanisms in the group environment. DOI 10.1093/mnras/stx973
  - Aim: To investigate how different quenching pathways correlate with morphology.
  - Dataset: GZ2 vote fractions (morphologies), SDSS/GALEX photometry.
  - Difference: Tabular data only approach, no imaging ML
- Géron et al. (2021): Galaxy Zoo: Stronger bars facilitate quenching in star forming galaxies. DOI 10.1093/mnras/stab2064
  - Aim: To prove that bars play a role in quenching star formation by funneling gas into central regions.
  - Dataset: DECaLS DR8 - https://www.legacysurvey.org/dr8/
  - Difference: Tabular only approach, No Hybrid Model Considered.

Our Contribution:

Unlike prior works that either:

1. use only image data (Dieleman, Hayat, Domínguez Sánchez), or
2. rely purely on tabular/statistical analyses (Smethurst, Géron),

our project directly compares image-only, tabular-only, and combined ML approaches to assess how different data modalities influence galaxy classification and quenching prediction.

Work Plan & Project Requirements

Gantt Chart

PROCESS	WEEKS											
	1	2	3	4	5	6	7	8	9	10	11	12
Datasets & Problem definition	■	■										
Data Preprocessing & EDA			■	■								
Training & Testing Baseline Models					■	■	■					
Training & Testing Neural Network							■	■	■	■		
Project Report & Code				■	■	■	■	■	■	■	■	
Project Presentation											■	■

Task Prioritization

TASK	PRIORITY
Selection of Datasets & Problem definition (R1) – *Mustansir, Vaishnavi, Adam, Janya, Ihsan*	Must
Data Preprocessing & Cleaning (R2) – *Vaishnavi, Adam, Mustansir*	Must
Exploratory Data Analysis (R2) – *Vaishnavi, Adam*	Must
Clustering Algorithm to find similarities (R2) – *N/A*	Wont
Application of at least 3 ML Algorithms (R3) – *Mustansir, Ihsan, Janya*	Must
Implementing Neural Network Models (R4) – *Vaishnavi, Mustansir*	Must
Performance Metrics and Analysis (R4) – *Adam, Janya, Ihsan*	Should
Model Fine-tuning and Optimization (R4) – *Adam, Janya, Ihsan*	Must
Reporting and Evaluation – *Mustansir, Vaishnavi, Adam, Janya, Ihsan*	Must

Thank you