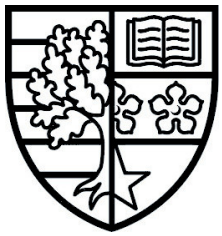
Galaxy Morphology and Quenching State Classification using ML models F20DL CW - 1 Report

by Group 10
Sri Sai Vaishnavi Chintha H00452920
Adam Aboushady H00385300
Janya Ratnakumar H00449225
Mustansir Eranpurwala H00432122
Mohammed Ihsan Fazal H00452069



Heriot-Watt University
School of Mathematical and Computer Sciences

GitHub repository link

20th November 2025

Table of Contents

1 Introduction

Understanding galaxy evolution, i.e. how galaxies form, transform, and eventually stop forming stars is one of the most fundamental areas of modern astrophysics. Two key observable properties crucial to this study are galaxy morphology (its structural shape, such as spiral or elliptical) and quenching state (whether it continues or has ceased forming stars).

Traditionally, projects like Galaxy Zoo have relied on human or crowd-sourced classification to label galaxy images, but the exponential growth of astronomical data from large surveys such as the Sloan Digital Sky Survey (SDSS) makes such manual efforts impractical. Recent research has demonstrated that deep learning models, particularly convolutional neural networks (CNNs), can achieve or surpass expert-level accuracy in classifying galaxy morphologies at scale [2, 3]. Studies comparing traditional algorithms (e.g., Decision Trees, Naïve Bayes) with deep learning architectures [1] show that CNNs offer superior precision for complex image-based tasks, while emerging self-supervised learning frameworks have demonstrated the ability to extract meaningful representations from unlabeled astronomical images, reducing dependence on costly manual labeling [5]. Furthermore, connections between a galaxy's morphology and its star formation history such as those revealed by the GaMorNet model [4] highlight the importance of analyzing both properties together

In this project, the aim is to build upon these advancements by applying and critically assessing a range of machine learning methods on astrophysical datasets from Galaxy Zoo 2 and SDSS. We are addressing two main research questions as follows:

(1) How do different types of data representations and machine learning models compare in their effectiveness for galaxy morphological classification?
(2) Can the galaxy quenching state be reliably classified using image-based and photometric feature-based deep learning approaches?

2 Dataset Description and Analysis

The tabular dataset used in this project is derived from the Galaxy Zoo 2 (Table 5) [7] public release, made available under an open-data, community-science license through the Zooniverse platform [8]. It contains 243,500 galaxies, each described by 233 tabular features, including object identifiers, sky coordinates, SDSS cross-match information (nominal/interval), total volunteer vote counts, weighted vote fractions, debiased vote fractions, and binary quality flags. The original dataset has a morphology string feature (gz2class), which we converted into a four-class morphology label: "elliptical", "spiral-unbarred", "spiral-barred", and "irregular". Summary statistics reveal uneven class distributions, with spirals dominating and irregulars underrepresented, as shown in Fig-1. Preprocessing involved dropping unused metadata fields and resolving missing and null values. To handle class imbalance, we downsampled all classes to the size of the smallest class, keeping the dataset balanced as shown in Fig-2. We also introduced an extra feature, "broad_class", which maps the morphology string into one of the four categories used for classification, and this feature was used as the final target class label for our models.

The image dataset used in this project originates from the SSL for Sky Surveys - SDSS Galaxy Datasets (Training dataset, redshift estimation) [6]. It contains 399,982 galaxy samples stored in an HDF5 file, where each sample consists of 107x107x5 photometric image representing the *u, g, i, r, and z* bands. Since the accompanying tabular dataset displayed an imbalance, the image dataset was matched with the tabular dataset using a common key

"specObjID" and then downsampled to maintain consistent counts. Preprocessing involved the extraction of the 5-band photometry to construct a new target class label "death_status", categorizing the galaxies as "quiescent","star-forming" or "transitional". The "broad_class" feature was also incorporated into this dataset to assist in the classification tasks as the target class.

3 Experimental Setup

The experimental setup evaluates the machine learning models employed for the classification of galaxy morphology and quenching states. The Baseline Machine Learning Algorithms implemented were Naive Bayes, k-Nearest Neighbours (k-NN), and Decision Trees. Naive Bayes provided a simple probabilistic baseline to assess whether morphology and quenching states can be classified using feature independence assumptions, while k-NN captured similarity-based patterns between galaxies. Decision trees were particularly useful as they handle mixed numerical and categorical Galaxy Zoo features without heavy preprocessing and provide feature-importance insights, making them a strong, interpretable baseline before more complex models. Neural network models included an MLP to model nonlinear relationships within tabular data and a CNN to directly learn spatial and spectral galaxy features from the SDSS image dataset.

Hyperparameter fine-tuning was performed across all models to identify optimal configurations. Naive Bayes was implemented using GaussianNB, with variance smoothing values tested between $10^{12}$and 10^{4} to assess numerical stability. For k-NN, several values of k (5, 11, 15, 19, 25) were explored to find the best neighborhood size. Decision trees were evaluated with different impurity criteria (gini vs entropy), maximum depths (5–30), and minimum samples per split (2–20). MLP experiments examined varying network depths, numbers of neurons (512-256-128, 256-128, and 128-64), and learning rates (0.001, 0.0001, and 0.0005), while the CNN was tuned for the number of convolutional layers and epochs. Performance across all models was assessed using standard classification metrics, including overall accuracy, precision, recall, F1-score, and confusion matrices. Preprocessing steps included standardization of tabular features to ensure stable and consistent training across all models.

4 Results

4.1 K Nearest-Neighbors

The KNN model achieved a train accuracy of 85.64% and test accuracy of 84.44%. The weighted averages for precision, recall, and F1-score were 0.8436, 0.8444, and 0.8417, respectively. These results indicate that KNN was able to capture the structure of the feature space reasonably well, though its performance was slightly lower than the Decision Tree model and comparable to Naive Bayes. The gap between train and test accuracy shows that the model generalised consistently without heavy overfitting. However, KNN can struggle when features aren't on the same scale or when classes overlap. Despite these limitations, the model demonstrated stable performance across all four morphological classes.

4.2 Naive Bayes

The Naive Bayes model achieved a train accuracy of 82.76% and test accuracy of 83.52%. The weighted averages for precision, recall, and F1-score were 0.85, 0.84, and 0.83, respectively.

These results indicate that the model performed reasonably well across all classes, though some misclassifications occurred due to the feature independence assumption inherent in Naive Bayes. Overall, it served as an effective baseline for comparison with more complex models.

4.3 Decision Tree

The Decision tree model achieved a train accuracy of 98.87% and test accuracy of 98.60%. The weighted averages for precision, recall, and F1-score were 0.99, 0.99, and 0.99, respectively. These results show that the Decision Tree learned the decision boundaries effectively, maintaining high generalisation performance with only a small gap between training and test accuracy. However, minor misclassifications are expected due to the model's tendency to create axis-aligned splits and its sensitivity to noisy or overlapping feature regions. Despite these limitations, the model remained highly interpretable and served as a strong benchmark for evaluating more advanced models.

4.4 Multi-Layer Perceptron

The MLP yielded very different results across the two tasks studied. The MLP model (Architecture: 512–256–128, LR = 1e-3) achieved a test accuracy of 47.27% and a weighted F1-score of 0.47. The performance is substantially lower compared to both the classical tabular models (Naive Bayes at 83.52% accuracy and Decision Tree at 98.60%) and the CNN, which achieved 64.67% accuracy. The overall results indicate that the photometric numeric features derived from the images do not have enough discriminative information that would allow for complex morphological classification.

In contrast, the MLP performed well for the classification of quenching states. For a small architecture (128–64, LR = 1e-3), the model achieved a test accuracy equal to 0.9934 with a low test loss of 0.0200, in addition to uniformly high values across all three classes, with precision, recall, and F1-scores always above 0.98. These results demonstrate that the photometric numeric features were strong indicators of the quenching state and that the MLP was able to effectively model relationships.

4.5 Convolutional Neural Network

The CNN achieved an accuracy of 64.67% and a weighted F1 Score of 0.63 using 2 Conv2D layers, 2 Maxpooling layers, 1 Flatten layer and 2 Dense layers. The performance is considerably lower than the tabular data models which had an average accuracy of 88.98% . This result indicates that image classification was less effective than feature models. This directly addresses the first research question by showing model performance varies across different data representation.This result also suggests that the quenching state isn't reliably predicted, meaning the original assumption that CNN could match or exceed the performance of tabular based models is not supported.

4.6 Comparision of models

The results from the morphological classification task provide a clear answer to the research questions:

RQ1: How do different types of data representations and machine learning models compare in their effectiveness for galaxy morphological classification?

Model	Test Accuracy (%)	Offset	Key Behaviour
Decision Tree	98.60	0.27	Best performer overall
KNN	84.44	1.20	Good balance, stable performance.
Naive Bayes	83.52	-0.76	Fastest but lowest accuracy among classical models.
MLP	47.27	1.86	Features insufficient, lowest accuracy
CNN	64.67	1.12	High cost, low accuracy

Table 1. Comparison of model performance on Morphology classification.

Model	Test Accuracy (%)	Offset	Key Behaviour
MLP	99.34%	0.59%	Near-perfect classification
CNN	77.96	1.22	less effective than the MLP

Table 2. Comparison of model performance on Quenching state prediction.

The comparison demonstrates that effectiveness varies drastically based on the data representation. The model trained on the tabular features, Decision Tree, achieved the highest test accuracy of 98.60%, establishing it as the most effective model. This high performance, combined with a minimal offset of 0.27. In contrast, the image-based deep learning approach (CNN) achieved only 64.67% accuracy. Despite its architectural complexity and higher computational cost, the CNN was significantly less effective than the simpler Decision Tree and even the KNN model with 84.44% accuracy. The poor performance of the MLP (47.27%) suggests that the photometric features do not possess the intricate, non-linear dependencies required for a deep learning model to accurately map them to morphology, contrasting with the Decision Tree's ability to exploit simpler, axis-aligned splits.

RQ2: Can galaxy quenching state be reliably classified using image-based and photometric feature-based deep learning approaches?

The performance of the MLP model on the tabular photometric features strongly indicates that galaxy quenching state can be reliably classified using this approach. The MLP achieved a near-perfect test accuracy of 99.34% with a minimal offset of 0.59, showing that the photometric features contain robust information for separating quenched and star-forming galaxies. This success is incredible, especially when compared with the same MLP's poor performance on the morphology task (47.27%). This confirms that the galaxy's quenching state is primarily encoded in its measurable photometric properties (like color and magnitude), which allows a relatively simple neural network to achieve maximum accuracy. In contrast, the CNN's 79.18% accuracy demonstrates that the image-based approach was significantly less effective and unsuitable for high-performance classification of this physical property. This result suggests that the visual features extracted by the CNN do not correlate strongly with the physical quenching mechanism.

5 Discussion

The results clearly demonstrate that the effectiveness of machine learning models for galaxy prediction strongly depends on both data representation and the target astrophysical property being classified. For morphology classification, feature-based traditional models such as Decision Trees and KNN significantly outperformed deep learning models, particularly those using raw image data. Decision Tree achieved the highest accuracy (98.60%) with minimal train-test gap, demonstrating that the Galaxy Zoo tabular features, especially debiased human vote fractions, contain rich, discriminative, and well-structured information that is highly suitable for rule-based decision-making. Naive Bayes and KNN also performed competitively (83–84% accuracy), suggesting that even with simple assumptions (feature independence or distance-based similarity), the tabular features encode morphology effectively. Conversely, the CNN model, despite using raw 107×107×5-band SDSS images, achieved only 64.67% accuracy, indicating that morphological cues such as bars, arms, and ellipticity are not easily captured by shallow CNNs without advanced architectures, rotational invariance, or larger pretraining datasets. The very poor performance of the photometric feature-based MLP (47.27%) further supports this, confirming that morphology is primarily defined by spatial structure, not by aggregated photometric values. In contrast, the classification of quenching states produced almost the opposite trend, highlighting how model suitability depends on the physical nature of the target property. The MLP trained on tabular photometric features achieved near-perfect test accuracy (99.34%) with high stability, demonstrating that quenching status is strongly encoded in color, luminosity, and magnitude-based measurements inline with astrophysical theory linking quenching to spectral energy distribution rather than galaxy shape. The CNN achieved moderate results (77.96%), showing that although spatial information provides some signal, photometric features are far more informative for quenching prediction than raw pixel data. This contrast highlights an important concept: morphology is primarily a structural property, making spatial image features more relevant (although not fully exploited by a simple CNN), while quenching is more closely linked to color and spectral properties, making tabular photometric features significantly more effective. These findings reinforce that the optimal choice of model depends on both the data representation and the physical nature of the target property, and confirms both research questions with strong experimental and theoretical support.

The practical implications of this work demonstrate that machine learning models could be developed for galaxy classification on a larger scale. However, the degree of realism and reliability varies across the different models and also depends on the data representation. The tabular models, particularly the Decision Tree and the MLP, showed the strongest and most stable performance, suggesting that feature-driven approaches could be realistically deployed in controlled pipelines where input data are well-processed and consistent. However, relying on real survey data could expose models to noise, missing photometric measurements, or unbalanced class distributions, therefore they ultimately would require additional robustness checks, such as cross-survey validation, uncertainty estimation, and automated data cleaning, for operational deployment. Furthermore, the requirement of structured, clean, and carefully curated input data means that real-world robustness would require ongoing recalibration and quality checks. In contrast, the CNN model's lower accuracy shows that image-based models, in their current form, are not yet reliable enough for scientific use. This is mainly due to limited architectural depth, reduced training data,

and the natural variability found in galaxy appearances. Meaningful improvements such as using deeper CNN architectures, training on larger and more diverse datasets, and applying stronger data augmentation would be required to address risks of overfitting and domain shift. Although the models do not pose any direct risks if deployed, the possibility of misclassifying rare galaxies highlights the need for uncertainty estimation, external benchmark testing, and human review before these systems can be integrated into real astrophysical workflows. Once refined, systems like these can meaningfully support researchers and industry practitioners by automating large scale morphological cataloguing, accelerating the analysis of massive surveys, and providing rapid, data-driven insights. This would hence reduce manual workload, improve consistency in classifications, and enable astronomers to focus on higher level scientific interpretation.

6 Conclusion

This project was conducted to evaluate how different machine learning algorithms perform when classifying the galaxies morphologies and their quenching state, which are two key factors in understanding galaxy evolution. The motivation for this project was the rapid growth in astronomical datasets, which makes manual classification like the ones conducted for Galaxy Zoo unfeasible and creates an opportunity for automated methods. The Galaxy Zoo 2 tabular dataset and the SDSS image dataset were used for this project since they cover a wide range of representations for the galaxies.

A set of algorithms were applied to these datasets to understand how model performance varies based on the datatype. Naive Bayes, k-NN and Decision Trees were implemented to assess the performance on the tabular data and provide a baseline for the neural networks. MLP and CNN were the neural networks of choice for this project, which were trained on the image data.

The analysis of our results answered both of our research questions clearly. The results showed that the models run on the tabular data, such as Decision Trees, performed strongly on the morphology classification, even outperforming the neural models, demonstrating that the features derived from human-consensus voting were the most viable discriminators for determining the complex structural properties of galaxies. Comparatively, the CNN performed notably worse which indicated that the dataset did not provide enough information for a more accurate morphology classification, as did the MLP with the numeric photometry features. For the quenching state, the MLP model, trained with extracted numerical photometric features (u, g, i, r, z bands), produced nearly perfect accuracy.

The overarching conclusion from this perspective is that the model selection must follow logically from the physical property being predicted: the type of galaxy morphology was best classified by structured, human-derived tabular features, and the evolutionary quenching state was best classified by photometric features employed in deep learning models. This hybrid approach offers the most reliable path for large-scale, automated galaxy classification in future astronomical surveys

References

[1] P. H. Barchi, R. R. de Carvalho, R. R. Rosa, R. A. Sautter, M. Soares-Santos, B. A. D. Marques, E. Clua, T. S. Gonçalves, C. de Sá-Freitas, and T. C. Moura. 2019. Machine and Deep Learning applied to galaxy morphology – A comparative study. *Astronomy and Computing* 28, 100334. doi:10.1016/j.ascom.2019.100334

[2] S. Dieleman, K. W. Willett, and J. Dambre. 2015. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly Notices of the Royal Astronomical Society* 450, 2, 1441–1459. doi:10.1093/mnras/stv632

[3] H. Domínguez Sánchez, M. Huertas-Company, M. Bernardi, D. Tuccillo, and J. L. Fischer. 2018. Improving galaxy morphologies for SDSS with Deep Learning. *Monthly Notices of the Royal Astronomical Society* 476, 3, 3661–3676. doi:10.1093/mnras/sty338

[4] et al. Ghosh. 2020. Galaxy Morphology Network: A Convolutional Neural Network Used to Study Morphology and Quenching in ~100,000 SDSS and ~20,000 CANDELS Galaxies. *<Add journal name here>* <Add volume>, <Add pages>. doi:<AddDOIifavailable>

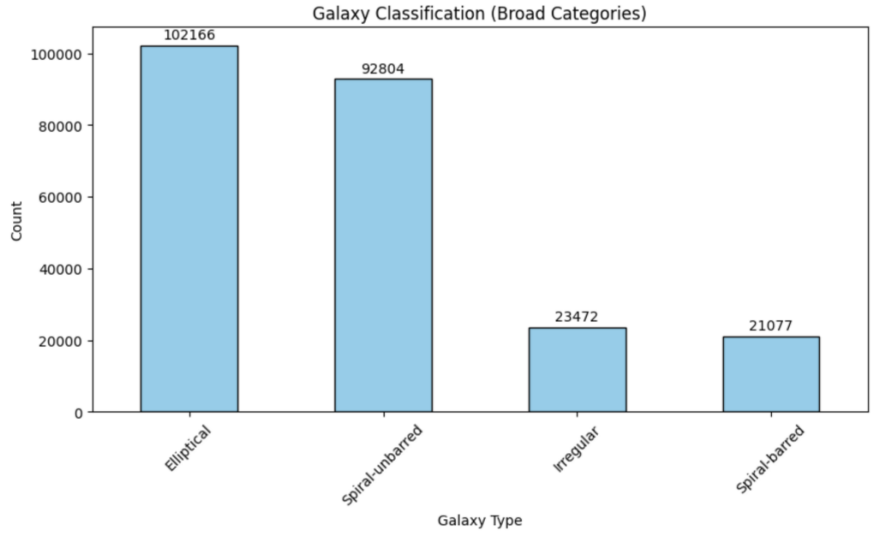
[5] M. A. Hayat, G. Stein, P. Harrington, Z. Lukić, and M. Mustafa. 2021. Self-supervised representation learning for astronomical images. *The Astrophysical Journal Letters* 911, 2, L33. doi:10.3847/2041-8213/abf2c7

[6] SSL for Sky Surveys. 2020. SDSS Galaxy Datasets: Training dataset, redshift estimation (HDF5 file sdss_w_specz_train.h5). https://portal.nersc.gov/project/dasrepo/self-supervised-learning-sdss/dataset.html. Dataset.

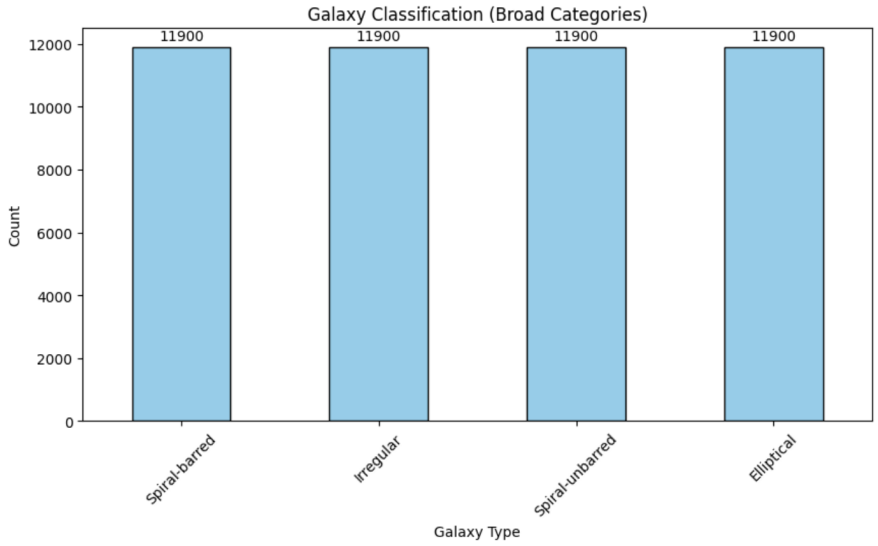
[7] Galaxy Zoo. 2013. Galaxy Zoo 2: Detailed morphological classifications for ~300,000 galaxies – Table 5. https://data.galaxyzoo.org/#section-7. Dataset.

[8] Zooniverse and Galaxy Zoo. [n. d.]. Galaxy Zoo – Data releases and catalogues. https://data.galaxyzoo.org/. Accessed: 2025-11-14.

A Appendix



(a) Class spread before normalization.



(b) Class spread after normalization.

B Results

Number of Neighbors	Test Accuracy
9	83.94%
13	84.10%
19	84.43%
21	84.44%
24	84.20%

Table 3. Table of KNN neighbours and test accuracy results

var_smoothing value	Test Accuracy
1e-4	78.31%
1e-6	83.52%
1e-8	83.24%
1e-10	83.93%
1e-12	82.70%

Table 4. Table of smoothing values and test accuracy results

criterion	maximum depth	minimum samples	test accuracy
gini	2	5	75.51%
entropy	2	5	75.87%
entropy	4	5	97.47%
entropy	4	2	97.47%
entropy	5	2	98.87%

Table 5. Table of smoothing values and test accuracy results

learning rate & neurons	Test Accuracy
256-128-64 & 1e-4	95.33%
256-128-64 & 5e-4	98.36%
512-256-128 & 1e-4	97%
512-256-128 & 5e-4	98.75%
128-64 & 1e-4	98.75%
128-64 & 5e-4	97%

Table 6. Table of MLP quenching state classification

learning rate & neurons	Test Accuracy
256-128-64 & 1e-4	45.53%
256-128-64 & 5e-4	48.05%
256-128-64 & 1e-3	48.30%
512-256-128 & 1e-4	46.87%
512-256-128 & 5e-4	49.00%
512-256-128 & 1e-3	43.57%
128-64 &1e-4	43.57%
128-64 & 5e-4	46.33%
128-64 & 1e-3	46.82%

Table 7. Table of MLP morphology state prediction

Epochs	**Layers**	**Performance**
5	2 Conv2D, 2 Maxpooling, 1 Flatten, 2 Dense	61.49%
5	3 Conv2D, 2 Maxpooling, 1 Flatten, 2 Dense	33.47%
10	2 Conv2D, 2 Maxpooling, 1 Flatten, 2 Dense	64.67%
10	3 Conv2D, 2 Maxpooling, 1 Flatten, 2 Dense	63.57%

Table 8. Table of CNN Morphology classification

C Contributions

Epochs	Layers	Performance
5	2 Conv2D, 2 Maxpooling, 1 Flatten, 2 Dense	77.46%
5	3 Conv2D, 2 Maxpooling, 1 Flatten, 2 Dense	75.42%
10	2 Conv2D, 2 Maxpooling, 1 Flatten, 2 Dense	76.17%
10	3 Conv2D, 2 Maxpooling, 1 Flatten, 2 Dense	74.23%

Table 9. Table of CNN Quenching state prediction

(1) Adam Aboushady (aha2003) - MLP model implementation for and report documentation
(2) Ihsan Fazal (mf2056) - Naive Bayes model implementation, Testing, GitHub environment setup, final documentation, and presentation
(3) Janya Rathnakumar (jr2068) - Decision Trees implementation, literature review, presentation, documentation
(4) Mustansir Eranpurwala (mte2000) - Image dataset feature extraction, data analysis, CNN model implementation, Documentation
(5) Vaishnavi Chintha (svc2000) - Tabular dataset preprocessing, data analysis and normalization, project formatting, and K-NN model implementation.

Fig. 2