Final Project: Statistical Modelling & Inference

Statistical Modelling of Galaxy Morphology

Albert Lamb Michael Duarte Gonçalves Tina Sikharulidze

December 2025

Abstract

What can a galaxy's light tell us about its shape, and can machines learn to see what humans see? This project investigates galaxy morphology using Sloan Digital Sky Survey (SDSS) and Galaxy Zoo 2 (GZ2) data through two complementary approaches. In Part 1, we model the relationship between photometric properties and human-perceived smoothness using Generalized Additive Model (GAM)s. Adaptive Least Absolute Shrinkage and Selection Operator (LASSO) reduces 36 candidate features to 15 interpretable predictors, and our GAM substantially outperforms linear baselines ($R^2 = 0.48$ vs. 0.33), revealing fundamentally non-linear relationships between physical measurements and visual perception. In Part 2, we develop an unsupervised pipeline combining Variational Autoencoder (VAE)s with Bayesian Gaussian Mixture Model (GMM)s to classify galaxies directly from images. The VAE compresses $128 \times 128 \times 3$ images into 16-dimensional latent representations, and Bayesian Information Criterion (BIC)-based model selection identifies $K = 4$ optimal clusters. Gibbs sampling provides full posterior uncertainty quantification. Together, these approaches demonstrate that interpretable statistical models and deep generative methods offer complementary insights into galaxy morphology.

Barcelona School of Economics

1 Introduction

Galaxy morphology—the visual classification of galaxies into types like spirals and ellipticals—is fundamental to modern astrophysics, as morphology is tightly linked to a galaxy's formation history. The massive volume of data produced by modern sky surveys, such as the SDSS, has far exceeded the capacity for manual classification by astronomers. This necessity for automation, and the desire to understand the physical drivers of visual appearance, is the primary motivation for this project.

The overall work is divided into two distinct parts: a Regression Analysis using quantitative metadata and an Unsupervised Learning approach using raw image data.

Part 1: Regression Analysis (Interpreting Human Judgment). The initial goal is to model and explain how a galaxy's quantitative physical properties relate to its visual morphology, specifically, smoothness of galaxies. The input to our model consists of quantitative SDSS features that describe the galaxies' brightness, size, and structure. These include photometric magnitudes across five optical filters (u, g, r, i, z), structural measurements such as the Petrosian radii, surface brightness, and the fraction of light fitted by different light profiles, as well as redshift values. The output variable is the debiased probability that a galaxy appears smooth, obtained from the GZ2 visual classification project. This measure represents collective human judgment, corrected for observational biases. We first employ adaptive LASSO regression to identify which features are most important, then explore a GAM with L2 regularization to output a predicted smoothness probability (P_{smooth}).

Part 2: Unsupervised Learning (Automating Classification). We transition to an unsupervised approach that utilizes raw galaxy images, aiming to reduce dependence on manual human labor. The input is an Red, Green, Blue (RGB) image. We use a VAE to learn a low-dimensional, structured latent representation ($\mathbf{z}$) of the galaxy image, following the statistical principles of Variational Inference. Finally, we use a Bayesian Mixture of Gaussians (MoG) model to estimate cluster membership probabilities. The output is a set of probabilities $P(\text{Cluster}_K|\text{Image})$, corresponding to the final automated morphological grouping inherent in the visual data.

2

Related Work

Research on galaxy morphology spans three major methodological directions: human-driven classification, supervised learning for prediction, and unsupervised representation learning.

2.1 Human-Driven Classification

Although large-scale surveys are increasingly automated, human annotation remains foundational. Galaxy Zoo 2 provides high-quality morphological labels, serving as the ground truth for supervised models (Lintott et al., 2008). Manual inspection is still required both for constructing training datasets and for validating automated approaches.

2.2 Supervised Learning Approaches

Modern state-of-the-art systems for morphology prediction rely on supervised Convolutional Neural Network (CNN) and Deep Learning. The trained models achieve over 95% accuracy (Bom et al., 2023). Their strengths include scalability and predictive performance, but they lack physical interpretability.

2.3 Interpretable Statistical Models

GAMs, introduced and formalized by Wood (2017), provide a flexible but interpretable framework for modeling nonlinear relationships in astronomy while retaining smooth, additive structure. However, standard frequentist inference is known to behave poorly in semiparametric models. Härdle et al. (2004) developed a wild bootstrap for GAMs that yields robust uncertainty quantification. A third methodological challenge involves inference after model selection. Fithian et al. (2014) introduced the framework of selective inference to ensure statistical validity with model selection. Their work motivates the use of a holdout dataset to preserve valid inference after applying feature selection.

2.4 Unsupervised Representation Learning

Some research seeks to automate morphological discovery by learning structure directly from raw images, independent of human labels. VAEs learn low-dimensional latent representations of galaxy images, which can be further clustered using GMMs to reveal natural morphological groupings (Xu et al., 2023; Saxena, 2025). These methods reduce dependence on citizen scientists, avoid human label biases, and enable the discovery of new morphological classes, though challenges remain regarding cluster interpretability and physical consistency.

3

Dataset Description and Preprocessing

3.1 Part 1: Regression Dataset

We used the GZ2 dataset (Galaxy Zoo Team, 2025), which contains crowd-sourced galaxy morphology classification derived from visual inspection of SDSS images (Lintott et al., 2008). The original table includes $n = 243{,}500$ instances and $d = 64$ covariates. The preprocessing steps, implemented in Python, include: removal of identifiers and positional columns; extinction correction for magnitudes using SDSS extinction values (SDSS Collaboration, 2023); correction of PETROMAG error encodings; construction of color indices between adjacent bands; calculation of concentration indices from Petrosian radii; surface brightness computation; log-transformations of radii and flux-like features; and removal of extreme outliers based on 10×Interquartile Range (IQR) thresholds. The distribution of all features after the preprocessing is provided in Appendix 3.1.1

After preprocessing, the final dataset consists of 229,168 instances and 36 features. The data was partitioned into two main sets (40%/60%): $n_{\text{selection}} = 91{,}667$ and $n_{\text{inference}} = 137{,}501$. The inference set was further split into training and testing (70%/30%): $n_{\text{train}} = 96{,}251$ and $n_{\text{test}} = 41{,}250$.

Target Variable (Y): The target is the debiased probability of smoothness, logit-transformed for modeling.

Feature Variables (X): All 36 features were standardized prior to modeling. The 36 features fall into four categories:

Photometric magnitude features include brightness measurements such as PSFMAG_R, FIBERMAG_R, DEVMAG_R, EXPMAG_R, CMODELMAG_R, and extinction-corrected Petrosian magnitudes across all bands. These characterize overall brightness under different light-profile assumptions.

Structural features capture morphological shape and size, including concentration index (CONC_R), logarithmic Petrosian radii, surface brightness (MU50_R), and de Vaucouleurs fraction (FRACDEV_R). These distinguish bulge-dominated from disk-dominated systems.

Color and spectral features include rest-frame absolute magnitudes (PETROMAG_MU through PETROMAG_MZ), encoding galaxy color indices, plus REDSHIFT and REDSHIFTERR for distance information.

Error features include logarithmic measurement uncertainties, encoding heteroskedasticity in SDSS photometry.

3.1.1 Sample Data

Index	target	PSFMAG_R	FIBERMAG_R	DEVMAG_R	EXPMAG_R	FRACDEV_R	MU50_R	CMODELMAG_R	REDSHIFT	REDSHIFTERR	PETROMAG_MU	PETROMAG_MG	PETROMAG_MR	PETROMAG_MI	PETROMAG_MZ	PETROMAG_U_corr	PETROMAG_G_corr	PETROMAG_R_corr
76715.0	0.4	17.2	17.4	16.2	16.5	1.0	19.6	16.2	0.1	0.0	-20.0	-20.7	-21.0	-21.2	-21.3	17.2	16.4	16.1
160003.0	0.2	17.7	17.7	16.6	16.9	1.0	19.6	16.6	0.1	0.0	-17.8	-19.4	-20.1	-20.5	-20.8	19.1	17.5	16.7

Index	PETROMAG_I_corr	PETROMAG_Z_corr	CONC_R	LOG_PETROR50_R	LOG_PETROR90_R	LOG_PETROR50_R_KPC	LOG_PETROMAGERR_U	LOG_PETROMAGERR_G	LOG_PETROMAGERR_R	LOG_PETROMAGERR_I	LOG_PETROMAGERR_Z	LOG_PETROMAGERR_MU	LOG_PETROMAGERR_MG	LOG_PETROMAGERR_MR	LOG_PETROMAGERR_MI	LOG_PETROMAGERR_MZ	LOG_DEVMAGERR_R	LOG_EXPMAGERR_R	LOG_CMODELMAGERR_R
76715.0	15.8	15.8	2.8	1.0	1.8	1.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
160003.0	16.3	16.0	3.1	0.9	1.7	0.9	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Figure 1: Sample data for Part 1 showing the 36 features after preprocessing.

The correlation analysis revealed several groups of highly collinear variables. The correlation heatmap is provided in Appendix B.

3.2 Part 2: Image Dataset

Due to computational limitations, we downloaded only 14,060 SDSS raw image cutouts and merged them with GZ2 debiased vote fractions to subsample images with reliable class labels. We applied strict probability thresholds: an image was assigned to the *smooth* class only if $p_{\text{smooth}} > 0.7$ while $p_{\text{disk}} < 0.3$ and $p_{\text{star}} < 0.3$ (and vice versa for disk). Since this is our first trial with these methods and raw SDSS images are quite noisy, we focused on selecting high-quality, separable galaxies. The final sample consisted of 2,696 smooth and 4,808 disk galaxies ($N = 7{,}504$), each resized to $128 \times 128 \times 3 = 49{,}152$ features.

Although we used classifications for subsampling, our approach is fully unsupervised—labels were used only for validating clustering choices. All RGB images were normalized to $[0, 1]$ and split using a stratified 80%/20% split: $n_{\text{train}} = 6{,}003$ and $n_{\text{val}} = 1{,}501$. Figure 2 shows representative examples.

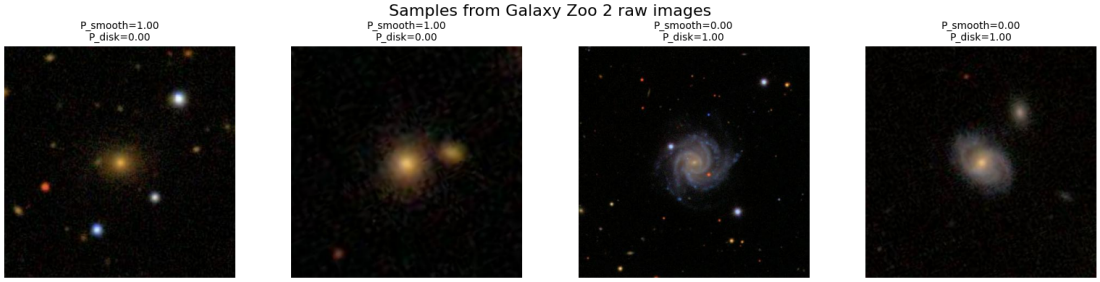

Figure 2: Example SDSS galaxy cutouts (left: smooth/elliptical, right: disk/spiral).

4

Regression Analysis

4.1 Baseline Analysis: Linear Model Performance and Feature Screening

We begin by establishing a performance baseline using linear regression models. Given our primary goal of interpretability, we started with Ordinary Least Squares (OLS). However, OLS provides no mechanism for feature selection, which is critical when interpreting 36 highly correlated features. Therefore, we shifted to penalized linear regression models. Regularization parameters λ were selected via 5-fold Cross-Validation (CV) to minimize test set mean squared error.

We employed three regularized regression techniques using the logit-transformed target Y_{logit}:

Ridge Regression (L_2 Penalty) adds a penalty term $\lambda \sum \beta_j^2$ that stabilizes coefficients by shrinking them towards zero, but retains all 36 features:

$$\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{d} \beta_j^2 \right\}$$

LASSO Regression (L_1 Penalty) adds a penalty $\lambda \sum |\beta_j|$ which shrinks coefficients exactly to zero, enabling automated feature selection:

$$\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{d} |\beta_j| \right\}$$

Adaptive LASSO refines the L_1 penalty with feature-specific weights $\hat{w}_j = 1/(|\hat{\beta}_{j,\text{Ridge}}|)^\gamma$, where $\gamma = 1$, penalizing features that were already small in the stable Ridge fit more aggressively:

$$\min_{\boldsymbol{\beta}} \left\{ \sum_{i=1}^{n} (y_i - \mathbf{x}_i^\top \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{d} \hat{w}_j |\beta_j| \right\}$$

Table 1: Comparison of Baseline Linear Models

Method	$\boldsymbol{\lambda}$	$\mathbf{R}^2_{\text{train}}$	$\mathbf{R}^2_{\text{test}}$	$\mathbf{corr}(y, \hat{y})$	$\mathbf{N}_{\text{features}}$
OLS	—	0.326	0.330	0.575	36
Ridge	0.039	0.326	0.330	0.575	36
Lasso (1)	0.0012	0.325	0.329	0.574	36
Lasso (2)	0.0085	0.300	0.303	0.551	18
Adaptive Lasso (1)	0.0008	0.325	0.329	0.574	35
Adaptive Lasso (2)	0.0304	0.2919	0.2942	0.5434	14

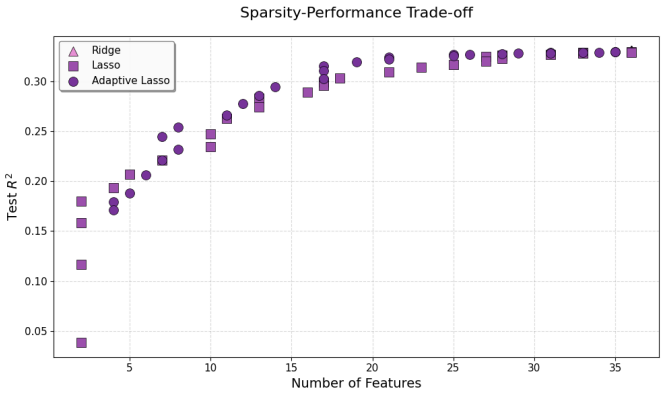

Figure 3: Sparsity-Performance Trade-Off for different methods

All three regularization methods follow remarkably similar trajectories, suggesting that the regularization strategy matters less than the number of informative features retained. The steep improvement between 1-15 features indicates that a core set of predictors captures most of the signal, while the plateau beyond 25 features suggests diminishing returns from additional complexity. Notably, Adaptive Lasso maintains competitive performance with fewer features in the mid-range (15-25 features), consistent with its theoretical advantage in feature selection, though the differences across methods remain modest in this application.

Key Findings: The comparable performance across all methods ($R^2_{\text{test}} \approx 0.33$) indicates that the predictive limitation stems from model bias rather than variance. The minimal Ridge regularization confirms that coefficient shrinkage provides no benefit, while LASSO and Adaptive LASSO reduce the feature space from 36 to approximately 25 features without sacrificing accuracy. This demonstrates that the linear assumption itself fundamentally limits model performance, motivating our transition to GAMs.

4.2 Generalized Additive Models

GAMs extend the linear framework by allowing each predictor to contribute through a smooth, nonlinear function while maintaining the additive structure that ensures interpretability. A GAM models the expected value of the response through a sum of smooth functions:

$$g(\mathbb{E}[Y]) = \beta_0 + \sum_{j=1}^{p} f_j(x_j)$$

where $g(\cdot)$ is a link function, and each f_j is a smooth function of x_j. These smooth functions are represented using penalized regression splines, which balance flexibility with smoothness. Each smooth function $f_j(x_j)$ is constructed as a linear combination of basis functions:$f_j(x_j) = \sum_{k=1}^{K_j} \beta_{jk} b_{jk}(x_j)$

where b_{jk} are basis functions (cubic splines in our case) and K_j is the number of basis functions (or "splines"). The model parameters $\{\beta_{jk}\}$ are estimated by minimizing a penalized likelihood:

$$\min_{\boldsymbol{\beta}} \left\{ -\mathcal{L}(\boldsymbol{\beta}) + \sum_{j=1}^{p} \lambda_j \int [f_j''(x)]^2 dx \right\}$$

where $\mathcal{L}(\boldsymbol{\beta})$ is the log-likelihood, and the penalty term $\lambda_j \int [f_j''(x)]^2 dx$ measures the wiggliness (curvature) of the smooth function f_j. The smoothing parameter λ_j controls the trade-off between fit and smoothness: larger λ_j produces smoother functions, while $\lambda_j \to 0$ allows more wiggly curves.

4.2.1 Feature Selection via Mutual Information

With 36 correlated features, direct interpretation of all GAM smooth functions would be challenging. We therefore employed Mutual Information (MI) to select the most informative features before fitting the GAM. MI measures the reduction in uncertainty about Y given knowledge of X: $I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$. Unlike linear correlation, MI captures non-linear dependencies and is therefore well-suited for identifying features that contribute non-linearly to the target. Within our feature selection set, we ranked all 36 features by their MI with the target and selected the top 15. The selected features are shown in Appendix C.

The MI-selected model achieves $R^2_{\text{test}} = 0.48$ (vs. 0.54 for the full 36-feature model). We accept this trade-off for interpretability.

4.2.2 Hyperparameter Tuning: Number of Splines and Regularization

We explored two key hyperparameters: the number of splines, and the smoothing parameter λ.

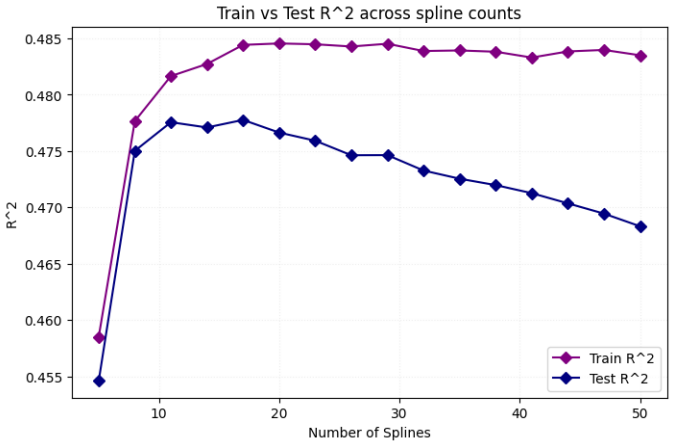

Figure 4: The effect of number of splines choice on R^2

Number of Splines (Figure 4): Figure 4 shows train and test R^2 as a function of the number of splines per feature. With very few splines (e.g., 5), the model is too rigid and underfits (low R^2). As

the number of splines increases beyond 10, train R^2 continues to rise slowly while test R^2 plateaus and then slightly declines after approximately 25 splines, indicating mild overfitting. The optimal range appears to be 10–25 splines, where the model is flexible enough to capture non-linearities but not so complex that it overfits.

Smoothing Parameter λ (Figure 5): Figure 5 displays a bubble chart where each bubble represents a combination of λ and number of splines, with bubble size proportional to test R^2. The key finding is that low values of λ (near zero) consistently yield the best performance, regardless of the number of splines. As λ increases, performance degrades, as the penalty forces the smooth functions to become increasingly linear, eventually recovering OLS-like behavior at very high λ. The insensitivity

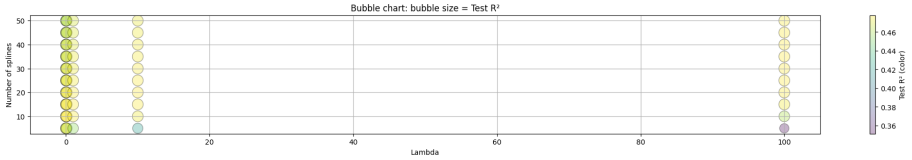

Figure 5: The effect of number of splines choice on R^2

to λ can be understood by examining the learned smooth functions. Most features exhibit naturally smooth, monotonic relationships with the target (e.g gentle sigmoids, exponential decays). These functions do not require aggressive smoothing penalties to avoid overfitting because the data itself does not support erratic, wiggly behavior. Even with $\lambda = 0$ (no penalty), the model learns smooth functions. Consequently, adding regularization is redundant: there are no oscillations to penalize. Figure 6 illustrates this by comparing GAM smooth functions at different λ values to linear regression slopes. At very low λ (e.g., $\lambda = 0.0001$), the GAM captures clear non-linearities (e.g., the sigmoid in PETROMAG_MI). At moderate λ (e.g., $\lambda = 1$), smooths become less flexible but still non-linear. At extremely high λ (e.g., $\lambda = 100{,}000$), the GAM smooth functions effectively recover linear regression.

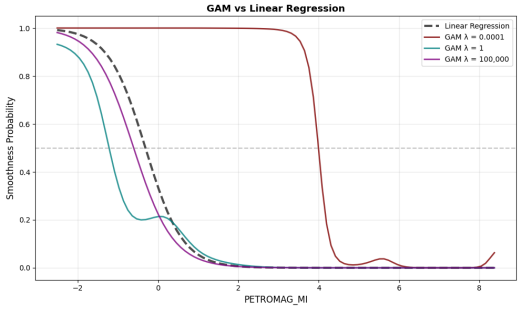

Figure 6: The effect of number of splines choice on R^2

4.2.3 Uncertainty Quantification: Wild Bootstrap Confidence Bands

The wild bootstrap (Härdle et al., 2004) resamples residuals while preserving the original predictor structure. For each bootstrap sample b, we: (1) compute fitted values $\hat{y}_i$ and residuals r_i; (2) generate random multipliers $v_i \in \{-1, +1\}$; (3) create bootstrap response $y_i^{(b)} = \hat{y}_i + v_i \cdot r_i$; (4) refit the GAM to obtain bootstrap smooth functions. Uniform 95% confidence bands are constructed from the empirical range of bootstrap curves.

Figure 7 displays the GAM partial dependence plots with confidence bands. Tight bands (e.g., LOG_PETROR50_R) indicate well-determined effects, while wider bands (e.g., PETROMAG_MZ) reflect data sparsity. Strong non-linearities are consistently captured across bootstrap samples, confirming these are genuine data features.

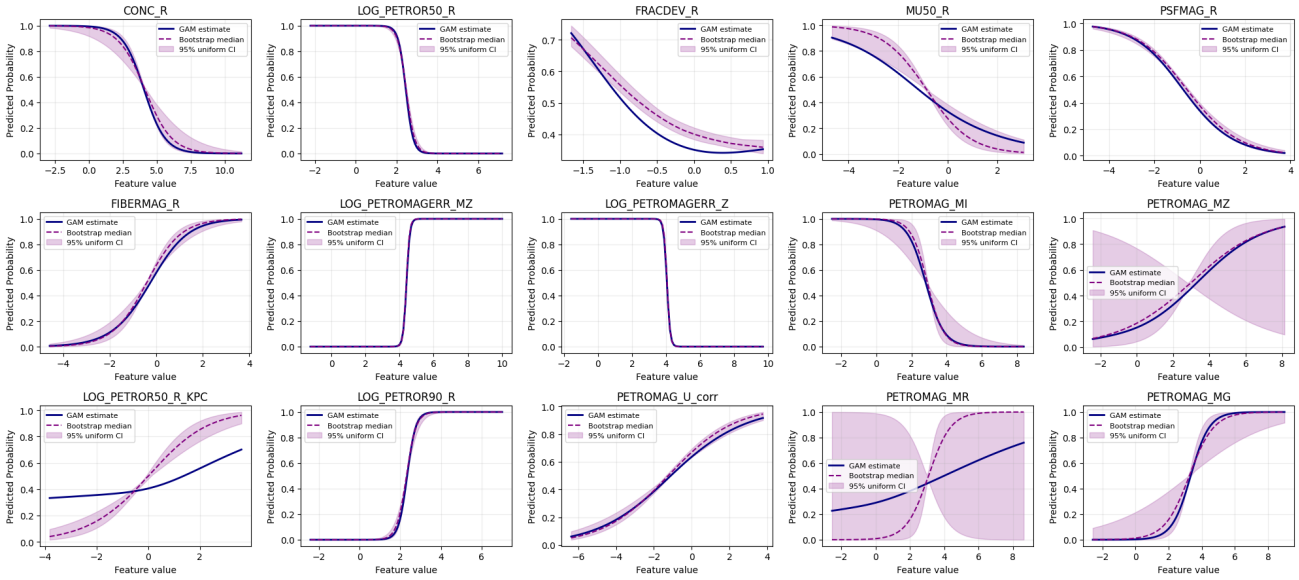

Figure 7: GAM partial dependence plots with wild bootstrap 95% confidence bands for each selected feature.

4.2.4 Statistical Significance and Effective Degrees of Freedom

Table 2 reports the significance and Effective Degrees of Freedom (EDF) of each smooth term. Out of 15 features, 14 are statistically significant at $p < 0.05$. The EDF values (ranging from 2.7 to 22.1) reflect varying degrees of non-linearity: high-complexity features (FRACDEV_R: 22.1, PETROMAG_MI: 19.1) show strong non-linear effects, while LOG_PETROMAGERR_Z (2.7) is nearly linear. The total EDF of 233.1 (65% of maximum) confirms the data supports substantial non-linearity.

Table 2: Statistical Significance and EDF of GAM Smooth Terms

Feature	EDF	p-value	Feature	EDF	p-value
CONC_R	12.7	< 0.001	PETROMAG_MI	19.1	< 0.001
LOG_PETROR50_R	18.0	< 0.001	PETROMAG_MZ	13.1	< 0.001
FRACDEV_R	22.1	< 0.001	LOG_PETROR50_R_KPC	18.7	0.219
MU50_R	18.4	< 0.001	LOG_PETROR90_R	15.8	< 0.001
PSFMAG_R	18.9	< 0.001	PETROMAG_U_corr	17.1	< 0.001
FIBERMAG_R	18.7	< 0.001	PETROMAG_MR	11.2	< 0.001
LOG_PETROMAGERR_MZ	18.7	< 0.001	PETROMAG_MG	9.0	< 0.001
LOG_PETROMAGERR_Z	2.7	< 0.001			

The model calibration plots can be seen in Appendix D.

4.3 Summary and Interpretation

The GAM achieves $R^2 = 0.48$ with 15 features, substantially improving over linear models ($R^2 \approx 0.33$). Key findings:

1. Non-linearity is essential: The true relationships between photometric features and the target are strongly non-linear.
2. Robust feature effects: Tight bootstrap confidence bands confirm that learned non-linearities are genuine data features.
3. Measurement error drives classification confidence: Two variables, LOG_PETROMAGERR_MZ and LOG_PETROMAGERR_Z, exhibit steep sigmoid transitions—objects with high uncertainty receive low classification probabilities.
4. Morphological features show gradual transitions: Structural parameters display smooth monotonic relationships, encoding real astrophysical differences.
5. Color features show threshold behavior: Features like PETROMAG_MR exhibit S-shaped curves consistent with known galaxy bimodality (red sequence vs. blue cloud).

5

UNSUPERVISED LEARNING: VAE AND BAYESIAN GMM

Our unsupervised pipeline operates in two stages: a VAE learns compact representations of galaxy images, then a Bayesian GMM discovers natural groupings within this latent space. The key insight is that if the VAE captures meaningful visual features, galaxies with similar morphologies should cluster together—even without explicit labels during training.

5.1 STAGE 1: VARIATIONAL AUTOENCODER

We assume each galaxy image $\mathbf{x} \in \mathbb{R}^D$ is generated from a latent vector $\mathbf{z} \in \mathbb{R}^d$ via:

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z} \mid \mathbf{0}, \mathbf{I})$$
$$p_\theta(\mathbf{x} \mid \mathbf{z}) = \mathcal{N}(\mathbf{x} \mid \mathbf{f}_\theta(\mathbf{z}), \sigma^2 \mathbf{I})$$

where $\mathbf{f}_\theta(\mathbf{z})$ is a deep convolutional neural network parameterized by θ. The true posterior $p(\mathbf{z}|\mathbf{x})$ is intractable since computing the posterior requires evaluating a high-dimensional integral with no closed-form solution. The likelihood is defined by a deep neural network, making the integral nonlinear and impossible to compute exactly. Following the variational inference framework, we approximate it with $q_\phi(\mathbf{z}|\mathbf{x})$ and maximize the Evidence Lower Bound (Evidence Lower Bound (ELBO)):

$$\log p(\mathbf{x}) \geq \mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z}))$$

We assume a diagonal Gaussian variational posterior:

$$q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}_\phi(\mathbf{x}), \text{diag}(\boldsymbol{\sigma}_\phi^2(\mathbf{x})))$$

where $\boldsymbol{\mu}_\phi(\mathbf{x})$ and $\boldsymbol{\sigma}_\phi(\mathbf{x})$ are outputs of the encoder. Further details are provided in Appendix E.

5.2 NETWORK ARCHITECTURE

We trained a VAE with the following architecture:

Encoder: The encoder network $q_\phi(\mathbf{z}|\mathbf{x})$ consists of 4 convolutional layers that progressively downsample the input: $3\times128\times128 \to 32\times64\times64 \to 64\times32\times32 \to 128\times16\times16 \to 256\times8\times8$. Each layer uses 4×4 kernels with stride 2, batch normalization, and ReLU activation. The final feature map is flattened (16,384 dimensions) and passed through two fully connected layers to produce the variational parameters $\boldsymbol{\mu}_\phi(\mathbf{x})$ and $\log \boldsymbol{\sigma}_\phi^2(\mathbf{x})$ for a 16-dimensional latent space.

Decoder: The decoder network $p_\theta(\mathbf{x}|\mathbf{z})$ mirrors the encoder architecture using transposed convolutions, with a final sigmoid activation to output pixel values in $[0, 1]$.

Training: The model was trained using the Adam optimizer with learning rate 5×10^{-4}, batch size 32, and $\beta = 0.5$ for 50 epochs with early stopping (patience=10). Gradient clipping (max norm 1.0) was applied to ensure training stability.

5.3 Stage 2: Bayesian Gaussian Mixture Model

Each latent vector is modeled as coming from one of K Gaussian components:

$$p(\mathbf{z}_n \mid \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}(\mathbf{z}_n \mid \boldsymbol{\mu}_k, \Sigma_k)$$

We place weakly informative conjugate priors: uniform Dirichlet on weights ($\alpha_k = 1$), wide Normal on means ($\mathbf{S}_0 = 10\mathbf{I}$), and minimum-information Inverse-Wishart on covariances ($\nu_0 = d + 2$).

Gibbs Sampling. We iteratively sample: (1) cluster assignments from $p(c_n = k \mid \mathbf{z}_n, \boldsymbol{\theta}) \propto \pi_k \mathcal{N}(\mathbf{z}_n \mid \boldsymbol{\mu}_k, \Sigma_k)$; (2) mixture weights from Dirichlet posteriors; (3) component means from Normal posteriors; (4) covariances from Inverse-Wishart posteriors. We initialize with k-means, run 5,000 iterations, and discard the first 1,000 as burn-in.

Model Selection. We select K minimizing $\text{BIC} = -2\log(L) + p\log(N)$.

5.4 Results

Latent Space Structure. Figure 9 shows the t-SNE projection of the VAE latent space. Even without clustering, smooth galaxies (high P_{smooth}) concentrate in distinct regions from disk galaxies, confirming the VAE captured morphologically meaningful features purely from reconstruction objectives.

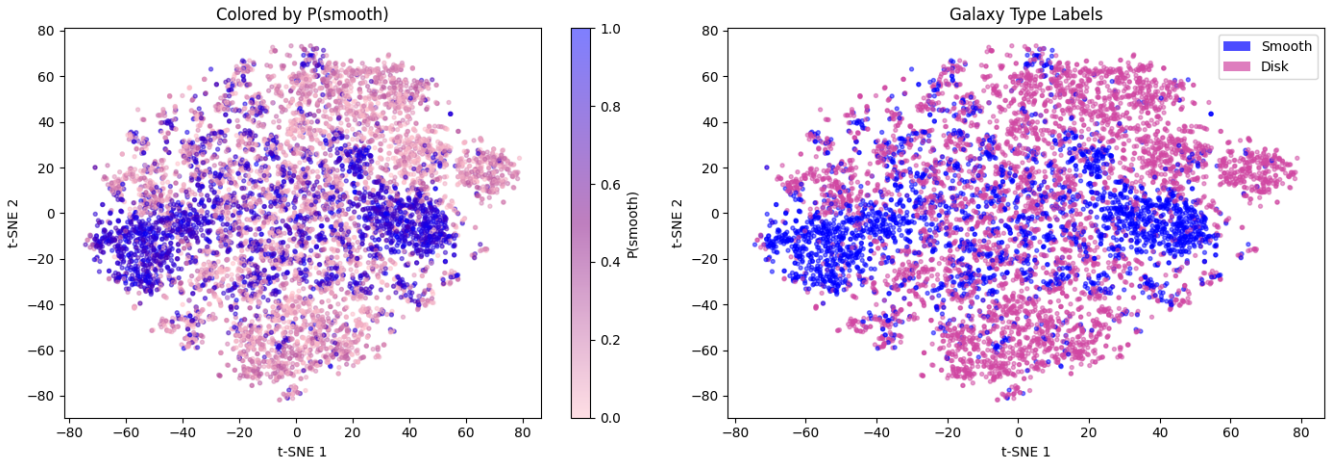

Figure 8: t-SNE projection of the 16-dimensional VAE latent space. Left: colored by continuous $P(\text{smooth})$. Right: colored by binary Galaxy Zoo classification.

The randomly selected reconstructions of images can be viewed in Appendix F.

Model Selection. Our analysis showed that $K = 4$ achieved the lowest BIC (177,231), indicating the best trade-off between fit and complexity. This suggests the data contains finer structure than the simple smooth/disk classification. Convergence diagnostics (Appendix G) confirm rapid convergence and stable mixing.

Cluster Structure. The four clusters show partial correspondence with smooth/disk classification but capture finer morphological distinctions. Posterior analysis yields well-identified mixture weights with tight credible intervals: $\pi_1 = 0.185 \pm 0.005$, $\pi_2 = 0.402 \pm 0.006$, $\pi_3 = 0.283 \pm 0.006$, $\pi_4 = 0.131 \pm 0.004$ (*cf.* Appendix G).

Although we filtered the Galaxy Zoo dataset for 2 primary morphological classes (smooth, disk), the GMM identified four distinct clusters in the learned latent space. In order to visually inspect the images, we randomly sampled 10 images out of each sample in figure. The VAE learned physically meaningful features beyond the smooth/disk classification present in the ground truth labels. Specifically, the model discovered that viewing angle plays a critical role in galaxy appearance, separating disk galaxies into edge-on and face-on orientations. Cluster 0 corresponds to elliptical and smooth galaxies. Clusters 1 and 2 both contain spiral/disk galaxies but differ primarily in orientation: Cluster 1 captures edge-on spirals where the disk is viewed from the side, while Cluster 2 contains face-on spirals. Cluster 3 represents intermediate morphologies, galaxies at confusing viewing angles. This unsupervised subdivision of the disk class demonstrates that the VAE successfully encoded geometric properties such as aspect ratio and structural orientation without explicit supervision.

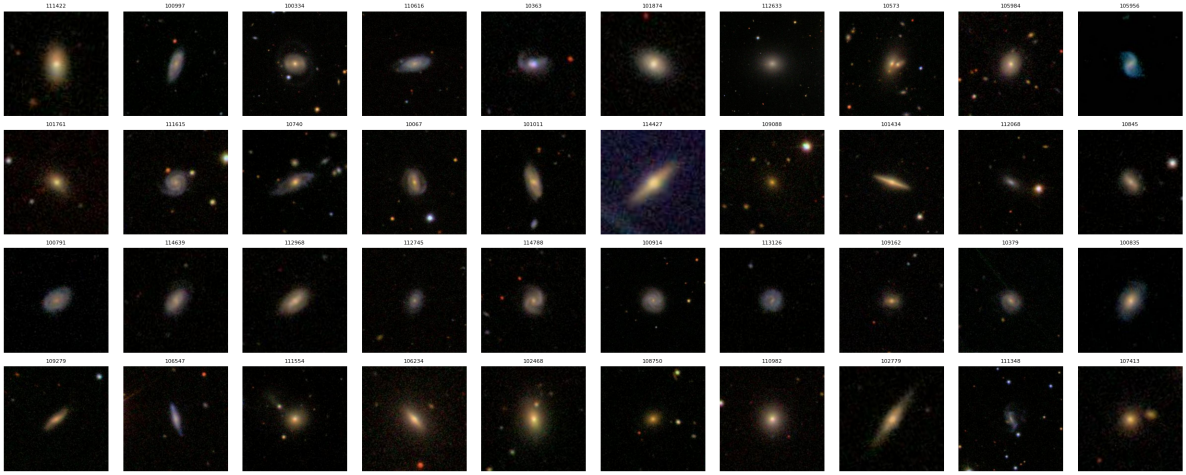

Figure 9: Randomly sampled images of galaxies for each cluster. Cluster labels were assigned based on maximum probability

6 Conclusion

This project investigated galaxy morphology through two complementary approaches.

The **regression analysis** demonstrated that GAMs substantially outperform linear models (R^2 = 0.48 vs. 0.33), revealing fundamentally non-linear relationships between photometric properties and human-perceived smoothness. MI-based feature selection reduced 36 features to 15 while maintaining interpretability, and wild bootstrap confidence bands confirmed that learned non-linearities are genuine data features. Measurement error features exhibit sharp threshold effects, while morphological and color features show gradual, physically interpretable relationships.

The **unsupervised pipeline** combining VAEs with Bayesian GMMs provides a principled framework for automated classification requiring no human labels during training. The VAE learned morphologically meaningful representations, BIC-based selection identified $K = 4$ optimal clusters capturing finer structure than binary labels, and Gibbs sampling provided full posterior uncertainty quantification with tight credible intervals.

Limitations and Future Work. The GAM explains only 48% of variance, suggesting additional features or interactions may be needed. The VAE-GMM pipeline's clusters don't perfectly correspond to morphological types, possibly capturing brightness or orientation variations. Future work could explore conditional generation from the fitted GMM, incorporate physical constraints into the VAE architecture, and investigate semi-supervised approaches that combine the interpretability of regression with the scalability of deep learning.

Acronyms

SDSS Sloan Digital Sky Survey

GZ2 Galaxy Zoo 2

OLS Ordinary Least Squares

LASSO Least Absolute Shrinkage and Selection Operator

GAM Generalized Additive Model

MI Mutual Information

EDF Effective Degrees of Freedom

IQR Interquartile Range

VAE Variational Autoencoder

GMM Gaussian Mixture Model

MoG Mixture of Gaussians

CNN Convolutional Neural Network

ELBO Evidence Lower Bound

KL Kullback–Leibler divergence

RGB Red, Green, Blue

CV Cross-Validation

BIC Bayesian Information Criterion

A
Feature Distributions

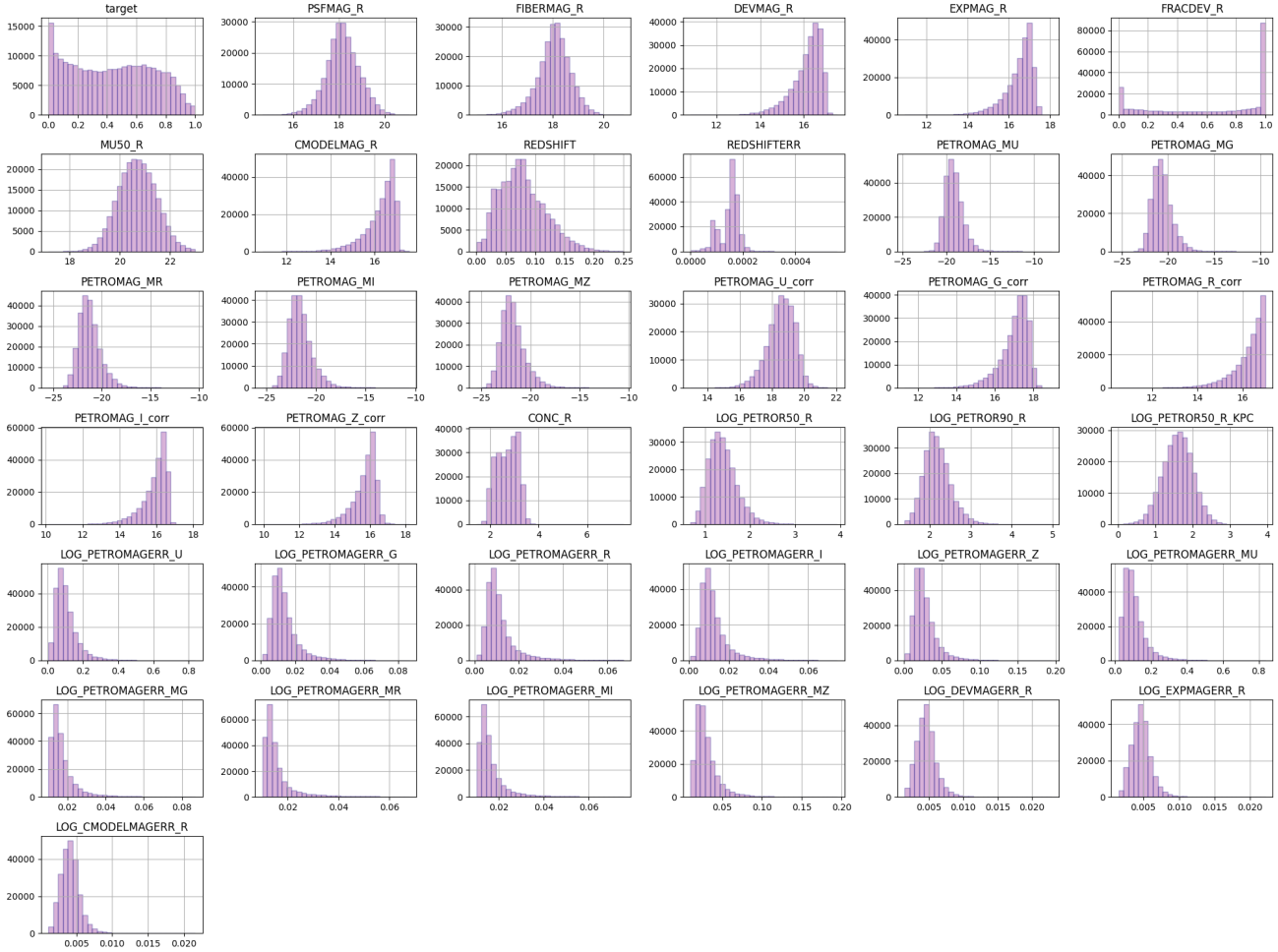

Figure 10: Histograms of target and all 36 features and after preprocessing.

B

Correlation Heatmap

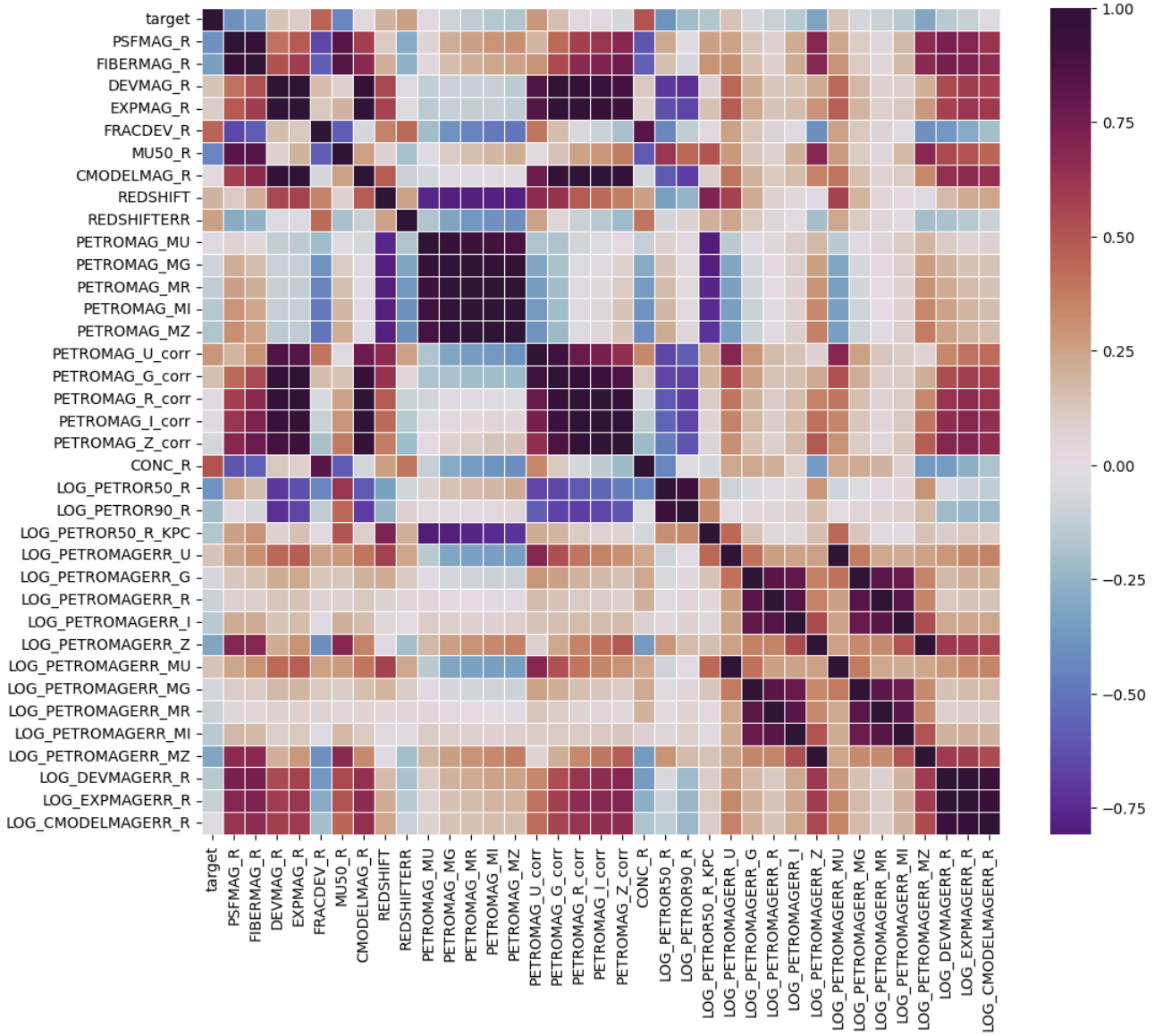

Figure 11: Correlation heatmap of the 36 features and target variable. Several groups of highly collinear variables are visible, primarily those measured across different photometric bands.

C

MI-Selected Features

Table 3: Features Selected via Mutual Information (Top 15)

Category	Selected Features
Structural	CONC_R, LOG_PETROR50_R, LOG_PETROR90_R LOG_PETROR50_R_KPC, FRACDEV_R, MU50_R
Photometric	PSFMAG_R, FIBERMAG_R, PETROMAG_U_corr
Color	PETROMAG_MI, PETROMAG_MZ PETROMAG_MR, PETROMAG_MG
Error	LOG_PETROMAGERR_MZ, LOG_PETROMAGERR_Z

D

Model Calibration

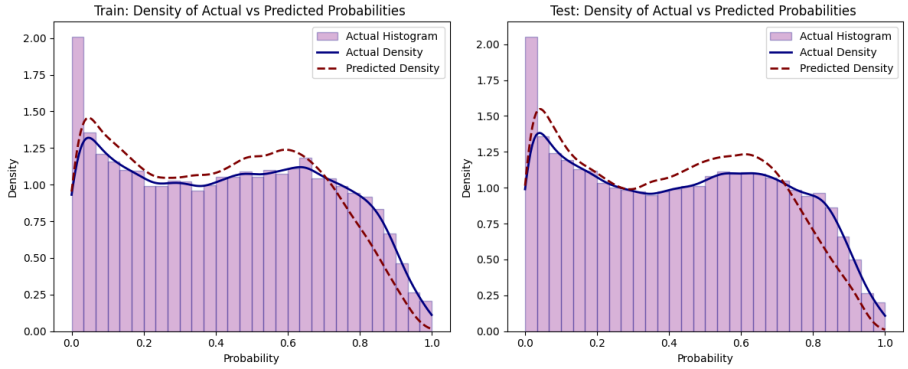

Figure 12: Density comparison of actual versus predicted smoothness probabilities. The bimodal predicted distribution suggests the model distinguishes two regimes (corresponding to different galaxy classes). Nearly identical train/test distributions confirm good generalization.

E

ELBO Decomposition

E.0.1 The Reparameterization Trick

Direct sampling from $q_\phi(\mathbf{z}|\mathbf{x})$ prevents gradient computation and therefore backpropagation, because sampling is not differentiable. We can apply the reparameterization trick and rewrite z as:

$$\mathbf{z} = \boldsymbol{\mu}_\phi(\mathbf{x}) + \boldsymbol{\sigma}_\phi(\mathbf{x}) \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$$

Now our parameters are not random and we can calculate their gradients. So now instead of the stochastic function, we have a deterministic function of external noise $\boldsymbol{\epsilon}$, enabling backpropagation through ϕ.

E.0.2 The VAE Loss Function

Training minimizes the negative ELBO, decomposed into two interpretable terms:

$$\mathcal{L}_{\text{VAE}}(\phi, \theta; \mathbf{x}) = \underbrace{\|\mathbf{x} - \mathbf{f}_\theta(\mathbf{z})\|^2}_{L_{\text{Recon}}\text{ (reconstruction)}} + \underbrace{\beta \cdot D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z}))}_{L_{\text{KL}}\text{ (regularization)}}$$

The first term is the deconstruction error, which measures how well the decoder reconstructs the input x from the sampled latent z. The second term measures how close the learned posterior is to the prior. If the posterior is very different from the prior, this term is large, penalizing the model. The hyperparameter controls the trade-off between these two terms.

KL Divergence for Diagonal Gaussian Posterior

The KL divergence between two Gaussians is:

$$D_{\text{KL}}(q_\phi(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})) = \frac{1}{2}\left[\log \frac{|\mathbf{I}|}{|\text{diag}(\boldsymbol{\sigma}^2)|} - d + \text{tr}(\mathbf{I}^{-1}\text{diag}(\boldsymbol{\sigma}^2)) + (\boldsymbol{\mu} - \mathbf{0})^\top \mathbf{I}^{-1}(\boldsymbol{\mu} - \mathbf{0}) \right]$$

Since $\mathbf{I}$ is the identity matrix and $\text{diag}(\boldsymbol{\sigma}^2)$ is diagonal, we have:

$$|\text{diag}(\boldsymbol{\sigma}^2)| = \prod_{j=1}^{d} \sigma_j^2, \quad \text{tr}(\text{diag}(\boldsymbol{\sigma}^2)) = \sum_{j=1}^{d} \sigma_j^2$$

Substituting, the KL divergence simplifies to:

$$\begin{aligned} D_{\mathrm{KL}}(q_\phi(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})) &= \frac{1}{2}\sum_{j=1}^{d}\Big[-\log\sigma_j^2 - 1 + \sigma_j^2 + \mu_j^2\Big] \\ &= \frac{1}{2}\sum_{j=1}^{d}\Big(\mu_j^2 + \sigma_j^2 - \log\sigma_j^2 - 1\Big) \end{aligned}$$

This is the closed-form expression used in VAE with diagonal Gaussian posteriors.

F

VAE Reconstructed Images

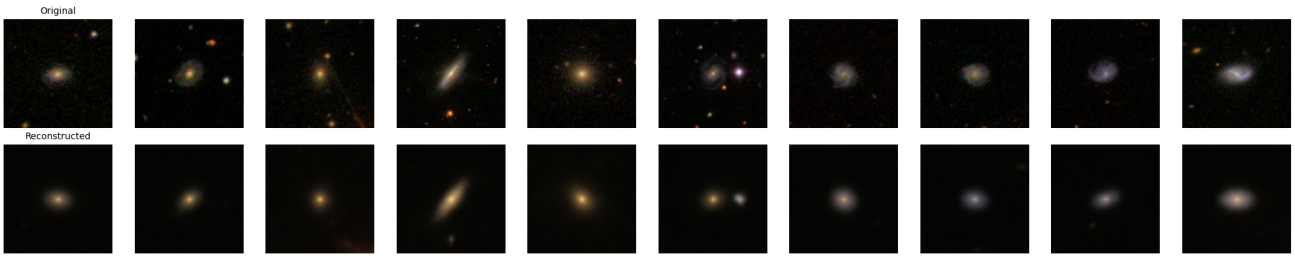

Figure 13: Representative original and reconstructed galaxy images from the test set. The VAE successfully captures the primary morphological features including galaxy orientation, central light, and overall light distribution. Fine details such as spiral structure, color gradients, and background stars are somewhat smoothed in the reconstructions, which is expected given the 16-dimensional latent representation. Notably, the model somewhat preserves the distinction between elliptical (round, smooth) and spiral (elongated, structured) galaxies, demonstrating that the latent representation encodes morphologically relevant information.

G
Convergence Diagnostics

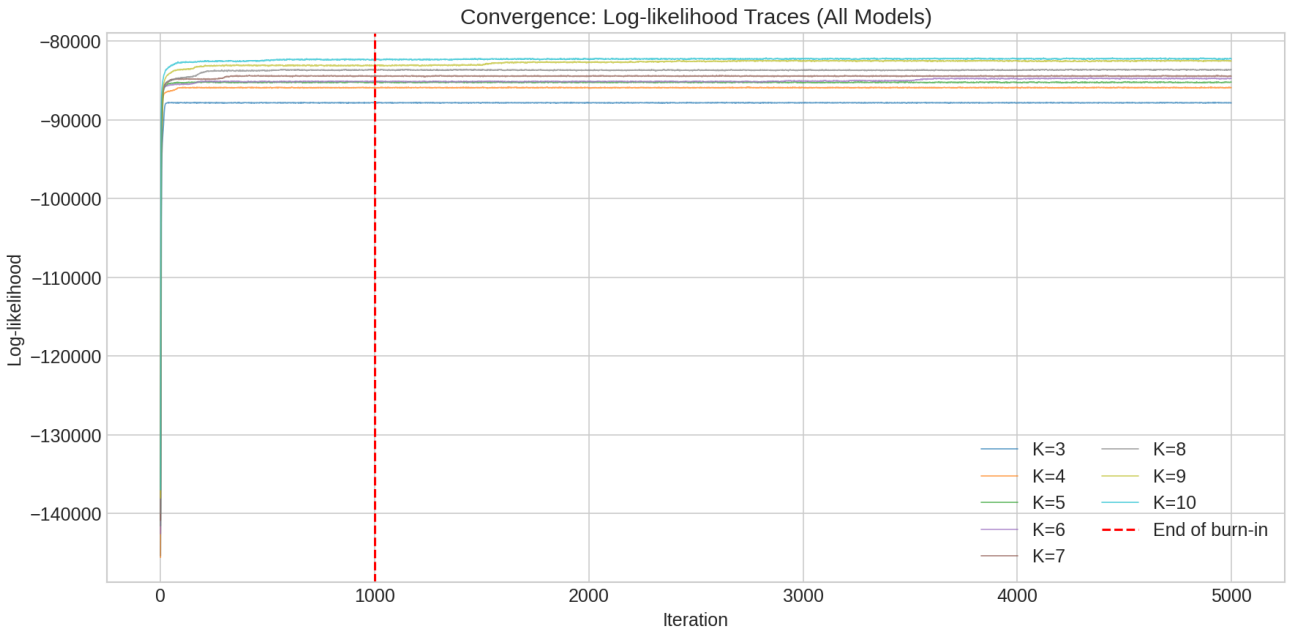

Figure 14: Log-likelihood traces for all GMM models ($K = 3, \ldots, 10$). All chains converge rapidly and show stable mixing after burn-in (red dashed line at iteration 1,000).

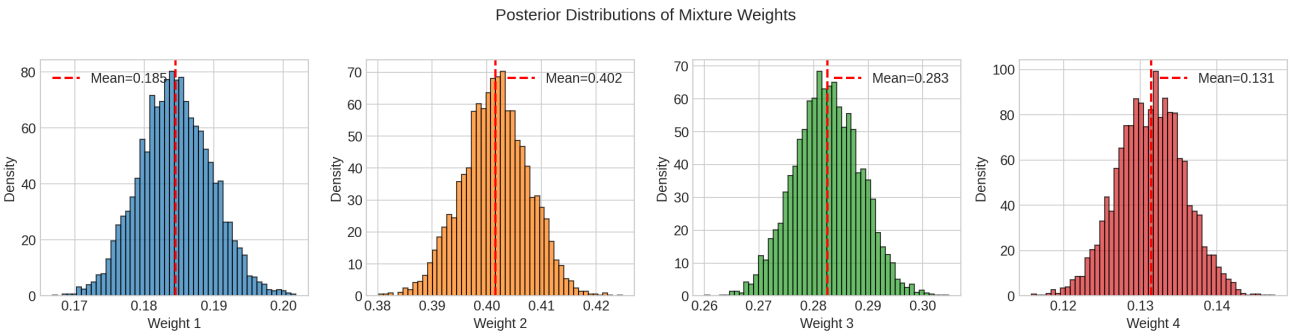

Figure 15: Posterior distributions of mixture weights for the $K = 4$ model. Tight, approximately Gaussian distributions indicate well-identified cluster proportions.

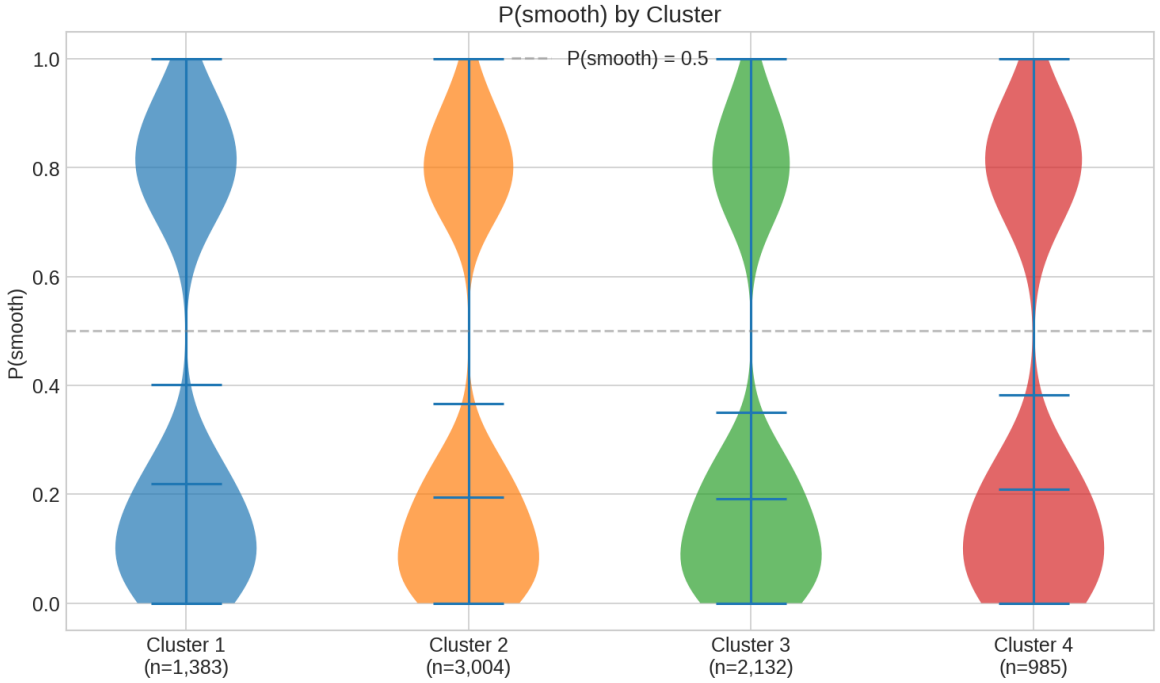

Figure 16: Distribution of $P(\text{smooth})$ within each cluster. All clusters contain both smooth and disk galaxies, suggesting the GMM captures morphological variation beyond binary classification.

H

Code Repository

All code used for data preprocessing, model training, and analysis is available in our GitHub repository:

GitHub Repository

Project Structure:

- `data/` — Raw datasets from Galaxy Zoo (not tracked due to capacity constraints)
- `preprocess.py` — Shared preprocessing utilities for all pipelines
- `01_Baseline/` — Regularized linear models (OLS, Ridge, Lasso, Adaptive Lasso)
- `02_GAM/` — Generalized Additive Models with spline transformations
- `03_VAE/` — Variational Autoencoder for latent representations from images
- `04_BMoG/` — Bayesian Mixture of Gaussians clustering via Gibbs sampling

References

Bom, C. R., Cortesi, A., Ribeiro, U., Dias, L. O., Kelkar, K., Smith Castelli, A. V., Santana-Silva, L., Lopes-Silva, V., Gonçalves, T. S., Abramo, L. R., Lima, E. V. R., Almeida-Fernandes, F., Espinosa, L., Li, L., Buzzo, M. L., Mendes de Oliveira, C., Sodré, L, J., Ferrari, F., Alvarez-Candal, A., Grossi, M., Telles, E., Torres-Flores, S., Werner, S. V., Kanaan, A., Ribeiro, T., and Schoenell, W. (2023). An Extended Catalogue of Galaxy Morphology using Deep Learning in Southern Photometric Local Universe Survey Data Release 3. *Monthly Notices of the Royal Astronomical Society*, 528(3):4188–4208.

Fithian, W., Sun, D., and Taylor, J. (2014). Optimal Inference after Model Selection. *arXiv preprint arXiv:1410.2597*.

Galaxy Zoo Team (2025). Galaxy Zoo data releases. `https://data.galaxyzoo.org/`. Public data releases from the Galaxy Zoo 1 & Galaxy Zoo 2 projects. Accessed: December 2025.

Härdle, W., Huet, S., Mammen, E., and Sperlich, S. (2004). Bootstrap Inference in Semiparametric Generalized Additive Models. *Econometric Theory*, 20(2):265–300.

Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, M. J., Nichol, R. C., Szalay, A., Andreescu, D., Murray, P., and Vandenberg, J. (2008). Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*, 389(3):1179–1189.

Saxena, A. (2025). Unsupervised Discovery of High-Redshift Galaxy Populations with Variational Autoencoders. *arXiv preprint arXiv:2511.05439*.

SDSS Collaboration (2023). Measures of flux and magnitude. `https://www.sdss.org/dr16/algorithms/magnitudes/`. SDSS documentation on photometric measurements. Accessed: December 2025.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2 edition.

Xu, Q., Shen, S., de Souza, R. S., Chen, M., Ye, R., She, Y., Chen, Z., Ishida, E. E. O., Krone-Martins, A., and Durgesh, R. (2023). From Images to Features: Unbiased Morphology Classification via Variational Auto-Encoders and Domain Adaptation. *Monthly Notices of the Royal Astronomical Society*, 526(4):6391–6400.