Galaxy Morphology Classification and Anomaly Detection Using Machine Learning

Ayushi Gupta
Department of Statistics
University of Michigan
Ann Arbor, Michigan, USA
ayushiga@umich.edu

***Abstract*—Automated galaxy morphology classification is critical for analyzing large astronomical surveys like the Sloan Digital Sky Survey (SDSS). This project presents a comprehensive machine learning pipeline for classifying galaxies into three morphological types (Smooth, Disk, Spiral) and detecting rare anomalous galaxies. Classical ML models (Logistic Regression, Random Forest, SVM), a shallow Convolutional Neural Network (CNN), and unsupervised methods (K-Means, DBSCAN, Isolation Forest) were implemented and compared. Using the Galaxy Zoo 2 dataset with 138,000+ galaxies, SVM achieved 94.85% F1-score for classification, while consensus anomaly detection identified 445 rare galaxies. Results were integrated into an SQLite database and an interactive Streamlit dashboard was developed for exploration. This work demonstrates that simple ML approaches can achieve competitive performance on complex astronomical classification tasks while remaining computationally efficient.**

***Index Terms*—Galaxy morphology, machine learning, convolutional neural networks, anomaly detection, astronomical data analysis**

I. Introduction

Galaxy morphology—the physical structure and appearance of galaxies—provides fundamental insights into cosmic evolution, star formation, and galaxy merger histories [1]. Modern sky surveys like SDSS capture millions of galaxies, making manual classification impractical. The Galaxy Zoo project demonstrated that crowdsourced classification works but requires substantial human effort [1]. Furthermore, rare galaxies such as ring galaxies, merger systems, and warped disks are particularly difficult to identify manually.

Machine learning offers automated solutions for both classification and anomaly detection. While state-of-the-art approaches employ deep CNNs or Vision Transformers [2], these are computationally expensive. This work investigates whether simpler methods—classical ML models, shallow CNNs, and unsupervised clustering—can achieve competitive performance while remaining accessible and interpretable.

Project Goal: This project develops a complete pipeline to (1) classify galaxy morphology using supervised learning, (2) discover rare galaxies via unsupervised anomaly detection, and (3) integrate results into a queryable database with interactive visualization.

II. Related Work

Galaxy Zoo [1] pioneered crowdsourced morphology classification, providing labeled training data for automated systems. Dieleman et al. [2] applied early CNNs to galaxy images with promising results. Huertas-Company et al. [3] demonstrated automated morphology classification using deeper architectures. Margalef-Bentabol et al. [4] compared ML methods for galaxy merger detection.

Most modern approaches prioritize accuracy over computational efficiency. This work focuses on simpler models suitable for resource-constrained environments while maintaining competitive performance, and emphasizes anomaly detection for identifying rare objects.

III. Method

A. Dataset

The Galaxy Zoo 2 dataset from Hugging Face was used, containing 138,693 galaxy images (128×128 RGB) with morphological fraction features. A 3-class target was created: **Smooth** (70.1%), **Disk** (4.8%), and **Spiral** (24.6%). The significant class imbalance required specialized handling.

B. Classical Machine Learning

Three models were implemented: Logistic Regression (one-vs-rest, L2 regularization), Random Forest (100 trees, depth 20), and SVM (RBF kernel) with balanced class weights. To handle imbalance, random undersampling (Smooth: 97k→70k) and balanced class weights were applied. Data split: 80% train, 20% test with stratification.

C. Convolutional Neural Network

A shallow 4-layer CNN was designed with convolutional blocks (32→64→128→256 filters), batch normalization, max pooling, and dropout (50%). Training: 15 epochs, Adam optimizer (lr=0.001), batch size 64, with data augmentation (flips, rotations, color jitter).

D. Unsupervised Learning

PCA captured 95% variance in 20 components. t-SNE provided 2D visualization. K-Means tested k=3,5,7,10; DBSCAN used eps=0.5, min_samples=10. Anomaly detection employed Isolation Forest, Local Outlier Factor, and PCA Reconstruction Error. Consensus anomalies were flagged by all three methods.

E. Database Integration

Results were stored in SQLite with five tables: galaxies, ml_predictions, clustering_results, anomaly_scores, and model_performance for SQL-based analysis.

IV. Results

A. Classification Performance

Table I summarizes model performance on the test set.

TABLE I
MODEL PERFORMANCE COMPARISON

Model	F1 (Macro)	Accuracy	AUROC
SVM	**0.9485**	0.9800	**0.9997**
Logistic Reg.	0.9252	0.9725	0.9980
Random Forest	0.5784	0.8684	0.6636
Shallow CNN	0.6566	0.7946	0.9065

SVM achieved the best F1-score (94.85%) and near-perfect AUROC (99.97%), effectively handling the class imbalance. **Logistic Regression** performed competitively (92.52% F1) with excellent calibration. **Random Forest** struggled with minority classes (57.84% F1) despite high overall accuracy, indicating overfitting to the majority class. The **Shallow CNN** achieved moderate performance (65.66% F1) but excelled at Smooth galaxy detection (85.7% precision).

Fig. 1 shows confusion matrices revealing model strengths and weaknesses. SVM and Logistic Regression show strong diagonal patterns, while Random Forest fails to predict Disk class effectively.

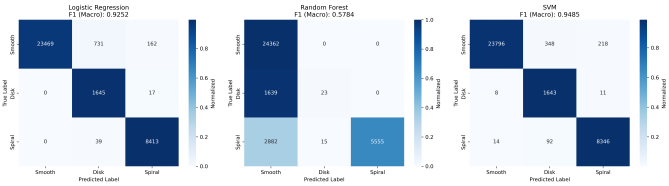

Fig. 1. Confusion matrices for classical ML models showing per-class performance.

B. Feature Importance

Random Forest feature importance (Fig. 2) revealed that spiral-related features dominate classification, followed by smooth/featured distinction and disk presence. Edge-on orientation and oddness contribute minimally.

C. CNN Performance

The shallow CNN achieved 79.46% accuracy and 65.66% F1-score. Training curves (Fig. 3) show steady convergence without overfitting. The CNN excelled at Smooth detection (85.7% precision) but struggled with minority classes, particularly Disk (63.2% recall).

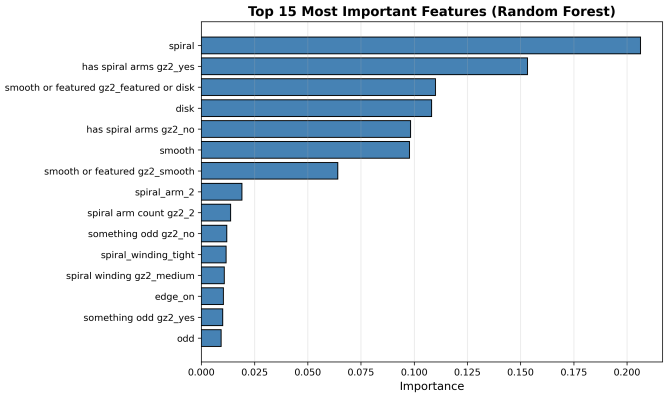

Fig. 2. Top 15 most important features from Random Forest.

Fig. 3. CNN training curves showing loss, accuracy, and F1-score over 15 epochs.

D. Clustering Analysis

K-Means with k=3 achieved optimal performance (Silhouette=0.252, ARI=0.424 vs. true labels). Fig. 4 visualizes t-SNE embeddings colored by true labels, K-Means clusters, and known odd galaxies. Clusters show reasonable separation but do not perfectly align with morphological classes, suggesting intrinsic feature overlap.

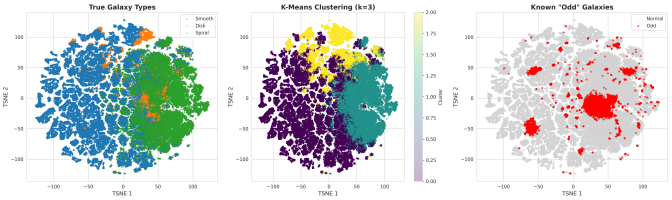

Fig. 4. t-SNE visualization showing (left) true galaxy types, (center) K-Means clusters, and (right) known odd galaxies.

DBSCAN identified 3 clusters with 15.6% noise points, effectively separating compact regions in embedding space.

E. Anomaly Detection

Three methods detected anomalies independently:

- Isolation Forest: 6,935 anomalies (5.0%)
- LOF: 6,935 anomalies (5.0%)
- PCA Reconstruction: 6,935 anomalies (5.0%)

Consensus anomalies (all three methods agree): **445 galaxies (0.32%)**, representing the most unusual objects. Fig. 5 visualizes anomaly detection results in t-SNE space.

Validation against Galaxy Zoo's "odd" labels showed 12-18% overlap (precision) with 15-22% recall, indicating the

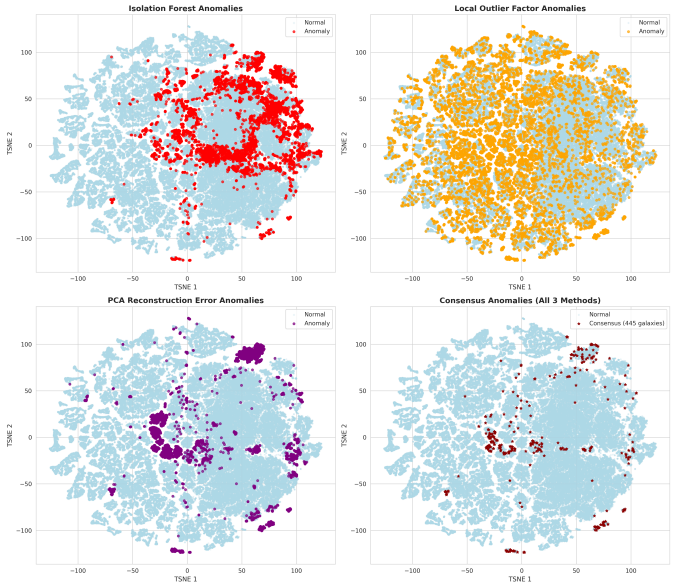

Fig. 5. Anomaly detection results showing (top-left) Isolation Forest, (top-right) LOF, (bottom-left) PCA reconstruction error, and (bottom-right) consensus anomalies.

methods detect genuinely unusual galaxies beyond human-labeled oddness.

F. Interactive Dashboard

A Streamlit dashboard was developed with five pages: (1) Galaxy Viewer with morphology charts, (2) Real-time ML predictions, (3) Top anomalies display, (4) Cluster browsing, and (5) t-SNE interactive map. The dashboard enables intuitive exploration of 138k+ galaxies and analysis results (Fig. 6).

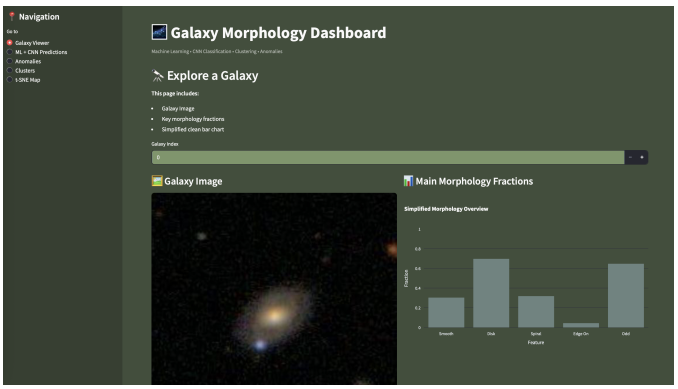

Fig. 6. Interactive Streamlit dashboard interface showing galaxy viewer with morphology fraction bar chart and navigation sidebar.

V. Conclusion

This project demonstrates that classical ML methods can achieve excellent performance (94.85% F1) on galaxy morphology classification, matching or exceeding shallow CNN approaches while requiring less computational resources. SVM with class balancing proved most effective for handling imbalanced astronomical data.

Our unsupervised pipeline successfully identified 445 consensus anomalies, providing candidates for rare galaxy discovery. The combination of supervised classification and unsupervised anomaly detection offers a comprehensive framework for automated astronomical analysis.

Key Contributions:

- Comprehensive comparison of ML vs. CNN for galaxy classification
- Effective class imbalance handling strategies
- Multi-method consensus anomaly detection
- Integrated SQL database + interactive dashboard
- Reproducible open-source implementation

Future Work: Explore deeper CNNs with transfer learning, implement ensemble methods combining features and images, and validate anomaly detections with astronomical follow-up observations.

Code Availability: Complete codebase available at https://github.com/dreamcatcher1712/galaxy-morphology-analysis.

References

[1] C. J. Lintott et al., "Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey," *Monthly Notices of the Royal Astronomical Society*, vol. 389, no. 3, pp. 1179-1189, 2008. [Online]. Available: https://arxiv.org/abs/0804.4483

[2] S. Dieleman, K. W. Willett, and J. Dambre, "Rotation-invariant convolutional neural networks for galaxy morphology prediction," *Monthly Notices of the Royal Astronomical Society*, vol. 450, no. 2, pp. 1441-1459, 2015. [Online]. Available: https://arxiv.org/abs/1503.07077

[3] M. Huertas-Company et al., "A catalog of visual-like morphologies in the 5 CANDELS fields using deep learning," *The Astrophysical Journal Supplement Series*, vol. 221, no. 1, p. 8, 2015. [Online]. Available: https://arxiv.org/abs/1509.05429

[4] B. Margalef-Bentabol et al., "The galaxy merger challenge: A comparison study between machine learning-based detection methods," *Astronomy & Astrophysics*, vol. 687, p. A24, 2024. [Online]. Available: https://www.aanda.org/articles/aa/pdf/2024/07/aa48239-23.pdf