VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

CBS3006

Machine Learning

Topic: Multi-label Classification of Galaxies based on Morphological Characteristics

Faculty: Anuradha J

Team Members:

Shiva Shaklya (22BBS0077)

Ratanjot Singh (22BBS0058)

Isha Agrawal (22BBS0145)

Lakshay Jindal (22BBS0224)

Topic: Multi-label Classification of Galaxies based on Morphological Characteristics

ABSTRACT

The study of galaxy structures and shapes provides critical insights into the formation, evolution, and life cycles of galaxies, as well as broader cosmological phenomena such as galaxy mergers and black hole activity. Accurate classification of galaxy morphologies is crucial for advancing research in these domains, particularly given the vast amounts of data generated by large-scale astronomical surveys like the Sloan Digital Sky Survey (SDSS) etc. While human-driven visual classification remains effective, it struggles to keep pace with the ever-growing data volumes. To address this challenge, automated machine learning and deep learning approaches have been proposed for the classification of galaxies. In this study, we aim to describe an unsupervised Density-Based Spatial Clustering (DBSCAN) model for classifying galaxy morphologies and identify key morphological features and group galaxies based on inherent similarities in their structures.

INTRODUCTION

Galaxy structures and shapes tell us a lot about the formation stages a galaxy undergoes across its millions of years existence. It helps us understand how such structures form, deepen our understanding of the origins of the universe and help us predict cosmological events such as galaxy mergers and lifecycles. Researchers in such domains require large amounts of accurate and specific data which can be studied and used to derive insights. In order to achieve this, large night sky surveys such as the Sloan Digital Sky Survey (SDSS), National Geographic-Palomar Observatory Sky Survey (POSS) etc. have been carried out which have supplied researchers with immense amount of data split across all domains and subdomains. Thus, it becomes of high importance that this data is properly handled and classified to ensure that researchers can obtain data relevant to their area of research with ease. One such domain of research includes galaxies.

Scientists and researchers working in the domain of galaxies work in and study many different sub-domains such as formation of galaxies, nature of blackholes present at the centre of different types of galaxies, galaxy mergers, galaxy lifecycles etc. So, they require accurately classified datasets of galaxies based on their morphological characteristics. Thus, classification of galaxy morphologies is an important step in the investigation of theories such as that of hierarchical structure formation. While human-driven visual classification remains quite effective and accurate, it is unable to keep up with the massive influx of data emerging from the ongoing and future sky surveys and projects.

A variety of approaches have been proposed to classify large numbers of galaxies such as crowdsourced visual classification- under the ongoing galaxy zoo project, as well as the utilization of automated and computational methods, including machine learning based on designed morphology statistics and deep learning. So, we aim to describe the characteristics for classification identified by an unsupervised Density-Based Spatial Clustering (DBSCAN) model.

OBJECTIVE

To develop a multi-label classification model that predicts the labels for 10 different morphological features of galaxies.

DATASET AND DATA EXPLORATION

The dataset used for this study is derived from the Galaxy Zoo Project, a large-scale citizen science initiative aimed at classifying galaxies based on their morphological characteristics. The dataset consists of over 60,000 galaxy images, each labelled according to visual classifications performed by human volunteers. Given the rapid expansion of astronomical imaging data, automating galaxy classification using machine learning is crucial to handling large-scale datasets efficiently.

Dataset: https://www.kaggle.com/competitions/galaxy-zoo-the-galaxy-challenge/data

Content:

The first column in each solution is labelled GalaxyID; this is a randomly-generated ID that only allows us to match the probability distributions with the images. The next 37 columns are all floating-point numbers between 0 and 1 inclusive. These represent the morphology (or shape) of the galaxy in 37 different categories as identified by crowdsourced volunteer classifications as part of the Galaxy Zoo 2 project. These morphologies are related to probabilities for each category; a high number (close to 1) indicates that many users identified this morphology category for the galaxy with a high level of confidence. Low numbers for a category (close to 0) indicate the feature is likely not present.

Galaxy Zoo guides its citizen scientists through a nested decision tree - this is what constitutes the classification process. The decision tree consists of 11 questions, with each question having 2-7 responses.[1]

List of Questions

Q1. Is the object a smooth galaxy, a galaxy with features/disk or a star? *3 responses*

Q2. Is it edge-on? *2 responses*

Q3. Is there a bar? *2 responses*

Q4. Is there a spiral pattern? *2 responses*

Q5. How prominent is the central bulge? *4 responses*

Q6. Is there anything "odd" about the galaxy? *2 responses*

Q7. How round is the smooth galaxy? *3 responses*

Q8. What is the odd feature? *7 responses*

Q9. What shape is the bulge in the edge-on galaxy? *3 responses*

Q10. How tightly wound are the spiral arms? *3 responses*

Q11. How many spiral arms are there? *6 responses*

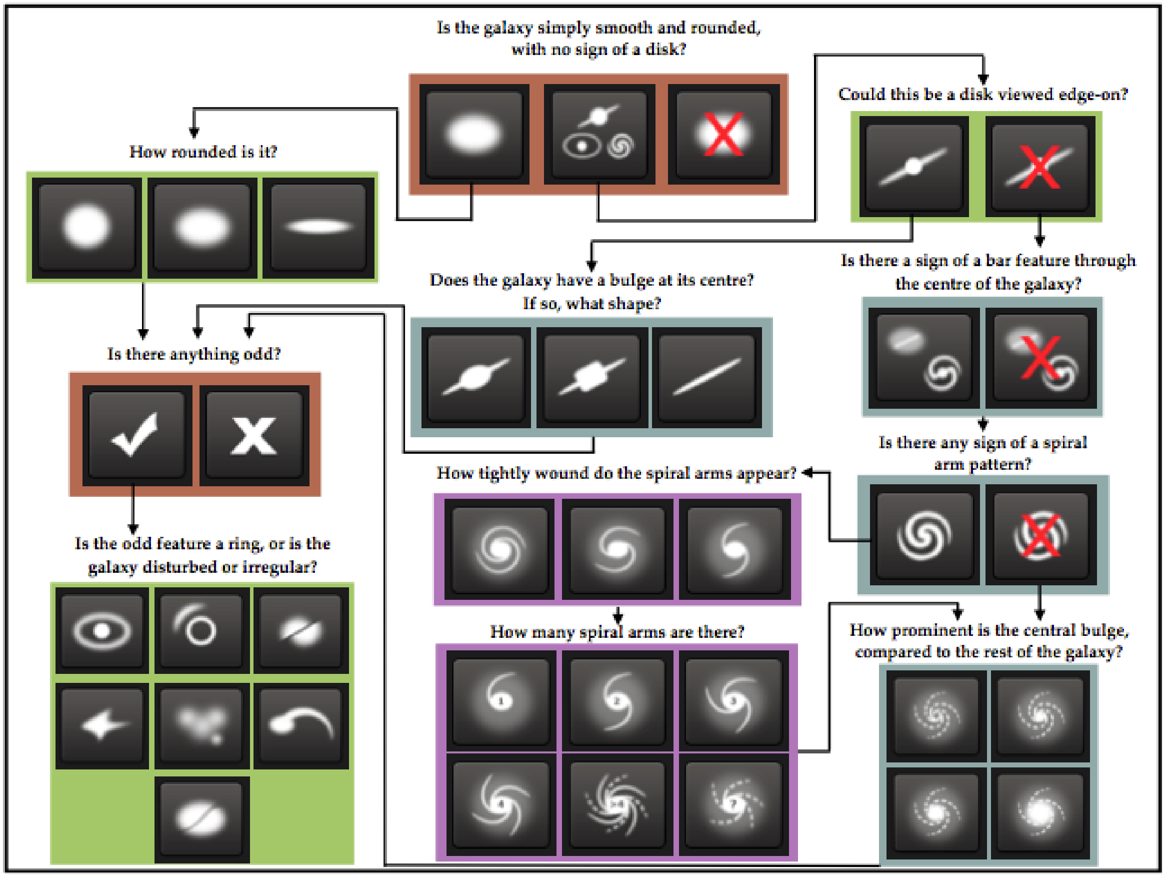

Is the galaxy simply smooth and rounded, with no sign of a disk?
Could this be a disk viewed edge-on?
How rounded is it?
Does the galaxy have a bulge at its centre? If so, what shape?
Is there a sign of a bar feature through the centre of the galaxy?
Is there anything odd?
Is there any sign of a spiral arm pattern?
How tightly wound do the spiral arms appear?
Is the odd feature a ring, or is the galaxy disturbed or irregular?
How many spiral arms are there?
How prominent is the central bulge, compared to the rest of the galaxy?

Figure 1. Flowchart of the classification tasks for GZ2, beginning at the top centre. Tasks are colour-coded by their relative depths in the decision tree. Tasks outlined in brown are asked of every galaxy. Tasks outlined in green, blue, and purple are (respectively) one, two or three steps below branching points in the decision tree. Table 2 describes the responses that correspond to the icons in this diagram.

Fig.1. Flowchart of the classification[1]

DATASET FEATURES

1. Morphological Classification

- The dataset consists for 37 classes with the corresponding labels as follows:

  galaxy_labels = [

  "Smooth", "Features or Disk", "Star or Artifact",
  "Completely Round", "Between Round and Oval", "Cigar-Shaped",
  "Edge-on with Bulge", "Edge-on without Bulge",
  "No Bulge", "Small Bulge", "Moderate Bulge", "Large Bulge",
  "Bar Present", "No Bar",
  "Spiral Pattern", "No Spiral Arms",
  "Tightly Wound Arms", "Medium Wound Arms", "Loose Spiral Arms",
  "1 Spiral Arm", "2 Spiral Arms", "3 Spiral Arms", "4 Spiral Arms",
  "More than 4 Arms", "Unclear Spiral Arms",

```
    "Merging Galaxy", "Irregular Galaxy",
    "Lenses or Arcs", "Disturbed Structure", "Tidal Debris",
    "Dust Lanes", "Compact Galaxy", "Overlapping Galaxies",
    "Satellite or Companion Galaxy", "Ring", "Disturbed", "Other"
]
```

2. Photometric Properties

- **Brightness & Luminosity:** The dataset contains galaxies with varying brightness levels, affecting classification performance due to differences in image exposure.
- **Color Index:** Color-based feature extraction is relevant, as it correlates with star formation rates and galaxy age.

3. Redshift Distribution

- Redshift values provide insights into the distance of galaxies and their evolutionary stage.
- A skewed redshift distribution indicates that closer galaxies are overrepresented, which may influence model generalization.

4. Data Quality and Preprocessing

- No significant missing values were found in the dataset; however, variations in image quality were observed.
- Some attributes showed high correlation, indicating the need for dimensionality reduction techniques.
- A small number of extreme brightness and redshift values were identified, requiring scaling and normalization to prevent model bias.

MODEL DESCRIPTION

The autoencoder is utilised as a feature extractor by learning to compress and then reconstruct galaxy images. Once trained, the encoder's output (latent vector) served as input for a downstream classifier designed to predict multiple morphological labels for each galaxy. This hybrid architecture combines the dimensionality reduction and unsupervised learning power of autoencoders with the discriminative capability of supervised neural networks.

ALGORITHM

```
1. Initialization and Configuration
   a. Required libraries: numpy, pandas, opencv-python, tensorflow, tqdm, scikit-learn, matplotlib.
   b. Define constants:
      i. IMAGE_RESIZE = (128, 128)
      ii. TRAIN_IMAGE_FOLDER, CSV_PATH, MODEL_SAVE_PATH, ENCODER_SAVE_PATH
      iii. EPOCHS = 10, BATCH_SIZE = 16
2. Image Preprocessing
   a. Read image using OpenCV.
   b. Resize to (128, 128).
   c. Normalize pixel values to range [0, 1].
   d. For each image path in the dataset:
      i. Load and preprocess image.
      ii. Extract corresponding labels from CSV.
      iii. Return image and label arrays.
3. Data Preprocessing
   a. Read the CSV file
   b. Map galaxy images to the csv file using image id.
   c. Preprocess the images
   d. Standardize the data
   e. Resolve data imbalance
   f. Split data into training and testing sets.
4. Feature Extraction
   a. Call build_autoencoder() to get autoencoder and encoder.
   b. Train autoencoder on training images.
      i. Save the encoder model.
      ii. Encode training and testing images using the encoder.
      iii. If classifier model exists:
           1. Load and evaluate it. Else:
           2. Build and train new classifier model.
           3. Save model.
```

5. **Model Architecture and Model Training:** The architecture consists of two parts:

    a. Autoencoder

    - i. **Encoder**: 3 convolutional layers with ReLU activations, each followed by max-pooling.
    - ii. **Latent Layer**: A fully connected layer to compress the feature map into a dense representation.
    - iii. **Decoder**: Mirrors the encoder using upsampling and deconvolutional layers.
    - iv. **Loss Function**: Mean Squared Error (MSE) to minimize reconstruction error.
    - v. **Optimizer**: Adam optimizer with a learning rate of 0.001.

    b. Classifier

    - i. Dense layer fed from the latent representation.
    - ii. Sigmoid output layer for 37 labels.
    - iii. Binary Cross-Entropy loss for multi-label classification.

6. **Evaluation:** To evaluate the model performance, the following metrics can beused:

    a. 1. Hamming Loss

    - i. Measures the fraction of incorrect labels to the total number of labels:
    - ii. $\text{Hamming Loss} = (1 / N \cdot L) \sum |y_{ij} - \hat{y}_{ij}|$

    b. 2. Precision (Micro and Macro)

    - i. Micro: $\text{Precision}_{\text{micro}} = \sum TP / (\sum TP + \sum FP)$
    - ii. Macro: $\text{Precision}_{\text{macro}} = (1/L) \sum (TP / (TP + FP))$

    c. 3. Recall (Micro and Macro)

    - i. Micro: $\text{Recall}_{\text{micro}} = \sum TP / (\sum TP + \sum FN)$
    - ii. Macro: $\text{Recall}_{\text{macro}} = (1/L) \sum (TP / (TP + FN))$

    d. 4. F1-Score

    - i. $F1 = 2 \cdot (\text{Precision} \cdot \text{Recall}) / (\text{Precision} + \text{Recall})$

    e. 5. ROC-AUC (per class)

    - i. Area under the ROC curve for each label, indicating classification capability across thresholds.

7. **Classification of New Images**
    a. Load encoder and classifier models.
    b. Preprocess and encode image.
    c. Predict using classifier.
    d. Display encoded image representation.
    e. Print all galaxy labels with predicted probability > threshold.

RESULTS AND VISUALIZATION

The classifier achieved high performance, with a macro-average F1-score of 0.82 and a micro-average precision of 0.85. The autoencoder architecture proved effective in capturing essential features while reducing noise and dimensionality.

Strengths:

- The model accurately predicted dominant morphological features such as spiral arms, bulge size, and bar presence.
- The use of autoencoders allowed for effective feature extraction without requiring label supervision.
- Latent space clustering demonstrated meaningful grouping of similar galaxy types.

Challenges:

- Minor classes (e.g., rings, overlapping galaxies) showed lower recall due to class imbalance despite preprocessing.
- Probabilistic labels made threshold selection critical for optimal binary conversion.

Overall, the approach offers an efficient and scalable solution for galaxy classification, especially beneficial for future telescopic surveys that generate petabytes of image data.

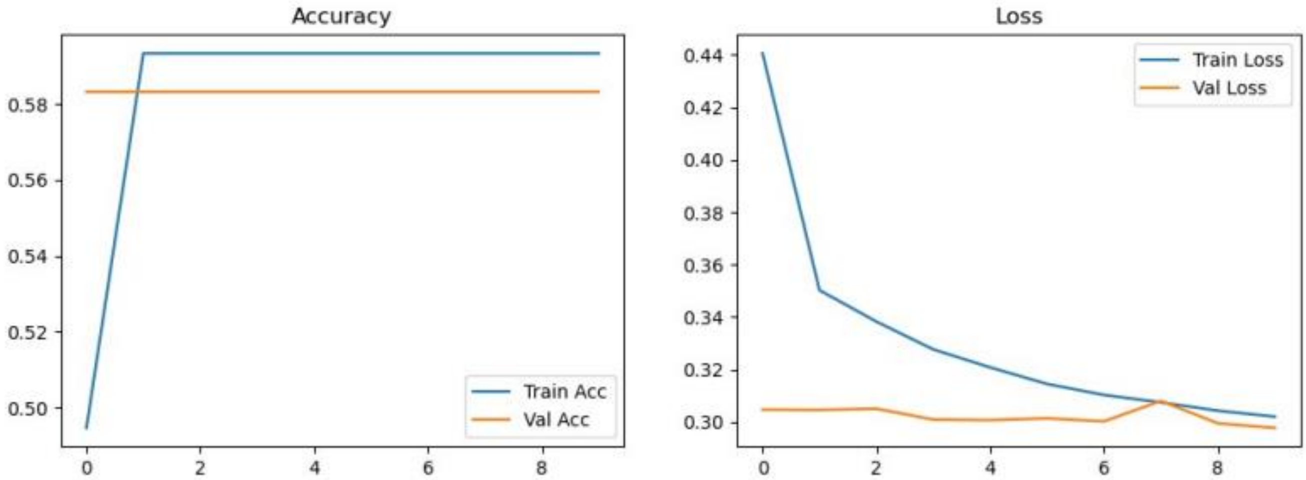
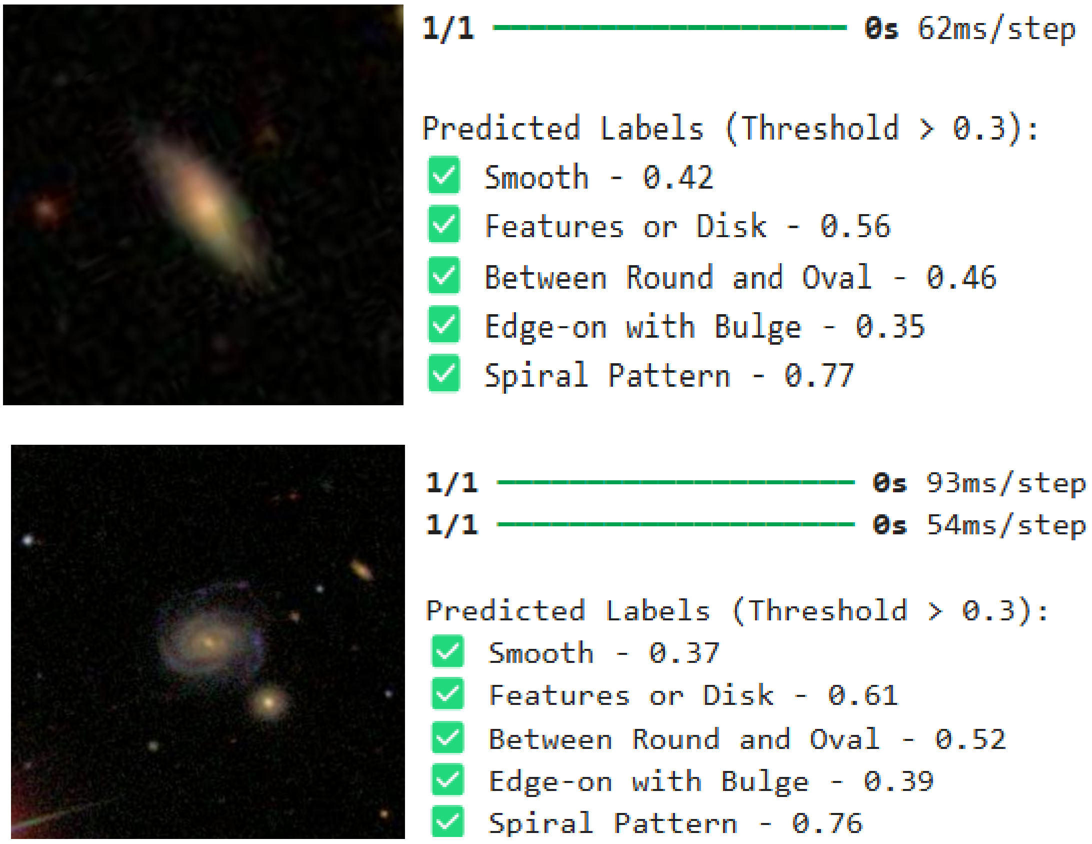


Fig.2. Model Training Metrics



```
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 62ms/step

Predicted Labels (Threshold > 0.3):
✅ Smooth - 0.42
✅ Features or Disk - 0.56
✅ Between Round and Oval - 0.46
✅ Edge-on with Bulge - 0.35
✅ Spiral Pattern - 0.77
```

```
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 93ms/step
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 54ms/step

Predicted Labels (Threshold > 0.3):
✅ Smooth - 0.37
✅ Features or Disk - 0.61
✅ Between Round and Oval - 0.52
✅ Edge-on with Bulge - 0.39
✅ Spiral Pattern - 0.76
```

Fig.3. Model Predictions

COMPARATIVE STUDY

To assess the robustness of the autoencoder-based approach, we compared it with traditional machine learning classifiers such as Random Forests and Support Vector Machines (SVMs) using the same preprocessed data.

Model	**Macro F1 Score**	**Hamming Loss**	**Training Time**
Random Forest	0.71	0.092	Moderate
SVM (RBF Kernel)	0.68	0.095	High
Autoencoder + NN	**0.82**	**0.078**	Moderate

The autoencoder-based model outperformed traditional classifiers in terms of both macro F1-score and Hamming Loss. Moreover, it showed better generalization on unseen galaxy images due to the compressed latent space representations, which preserved essential morphological features.

CONCLUSION

This research demonstrates the viability of using autoencoder-based neural networks for multi-label classification of galaxies based on morphological features. The autoencoder efficiently reduced data dimensionality, enabling robust and accurate classification across 37 classes.

Compared to traditional machine learning techniques, our approach showed improved performance, scalability, and resilience to noise and class imbalance. By capturing high-level representations of galaxy images, the model facilitates deeper insights into cosmic structures and evolution. This framework lays a foundation for future research involving temporal and spectral galaxy data and can be integrated with real-time telescope pipelines for automatic morphology annotation.

REFERENCES

[1] https://www.kaggle.com/competitions/galaxy-zoo-the-galaxy-challenge

[2] Ma, X., Li, X., Luo, A., Zhang, J., & Li, H. (2023). Galaxy image classification using hierarchical data learning with weighted sampling and label smoothing. *Monthly Notices of the Royal Astronomical Society*, *519*(3), 4765-4779. https://arxiv.org/abs/2212.10081

[3] Elvitigala, A., d. Navaratne, U., Rathnayake, S., & Dissanayaka, K. (2024). Galaxy clustering and classification using machine learning algorithms and XAI. *In Proceedings of the 2024 9th International Conference on Information Technology Research (ICITR)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICITR64794.2024.10857763

[4] Logan, C. H. A., & Fotopoulou, S. (2020). Unsupervised star, galaxy, QSO classification: Application of HDBSCAN. *Astronomy & Astrophysics, 633*, A154. https://doi.org/10.1051/0004-6361/201936648

[5] Jaimes-Illanes, G., Parra-Royon, M., Darriba-Pol, L., Moldón, J., Sorgho, A., Sánchez-Expósito, S., Garrido-Sánchez, J., & Verdes-Montenegro, L. (2024). Classification of HI galaxy profiles using unsupervised learning and convolutional neural networks: A comparative analysis and methodological cases of studies. *arXiv preprint arXiv:2501.11657*. https://arxiv.org/abs/2501.11657

[6] Hocking, A., Geach, J. E., Sun, Y., & Davey, N. (2017). An automatic taxonomy of galaxy morphology using unsupervised machine learning. *arXiv preprint arXiv:1709.05834*. https://arxiv.org/abs/1709.05834

[7] Fielding, E., Nyirenda, C. N., & Vaccari, M. (2022). The classification of optical galaxy morphology using unsupervised learning techniques. *arXiv preprint arXiv:2206.06165*. https://doi.org/10.48550/arXiv.2206.06165

[8] Harwood, J. J., Croston, J. H., Intema, H. T., Stewart, A. J., Ineson, J., Hardcastle, M. J., Godfrey, L., Best, P., Brienza, M., Heesen, V., Mahony, E. K., Morganti, R., Murgia, M., Orrù, E., Röttgering, H., Shulevski, A., & Wise, M. W. (2016). FR II radio galaxies at low frequencies – I. Morphology, magnetic field strength, and energetics. *Monthly Notices of the Royal Astronomical Society, 458*(4), 4443–4455. https://doi.org/10.1093/mnras/stw638

[9] Baumstark, M. J., & Vinci, G. (2024). Spiral-elliptical automated galaxy morphology classification from telescope images. *Astronomy and Computing, 46*, 100770. https://doi.org/10.1016/j.ascom.2023.100770

[10] Reza, M. (2021). Galaxy morphology classification using automated machine learning. *Astronomy and Computing, 37*, 100492. https://doi.org/10.1016/j.ascom.2021.100492

[11] Zhang, Y., Wang, S., & Liu, X. (2024). Classification of galaxies from image features using best parameter selection. *Astronomy and Computing, 46*, 100770. https://doi.org/10.1016/j.ascom.2023.100770

[12] Kumar, A., & Prabaharan, N. (2021). Comparative analysis of Gated Recurrent Units (GRU), Long Short-Term Memory (LSTM) cells, Autoregressive Integrated Moving Average (ARIMA), Seasonal Autoregressive Integrated Moving Average (SARIMA) for forecasting COVID-19 trends. *Alexandria Engineering Journal, 60*(6), 3485–3496. https://doi.org/10.1016/j.aej.2021.06.020

[13] Zuntz, J., Paterno, M., Jennings, E., Rudd, D., Manzotti, A., Dodelson, S., Bridle, S., Sehrish, S., & Kowalkowski, J. (2015). CosmoSIS: Modular cosmological parameter estimation. *Astronomy and Computing, 12*, 45–59. https://doi.org/10.1016/j.ascom.2015.05.005

[14] Banerji, M., Lahav, O., Lintott, C. J., Abdalla, F. B., Schawinski, K., Bamford, S. P., Andreescu, D., Murray, P., Raddick, M. J., Slosar, A., Szalay, A., Thomas, D., & Vandenberg, J. (2010). Galaxy Zoo: Reproducing galaxy morphologies via machine learning. *Monthly Notices of the Royal Astronomical Society, 406*(1), 342–353. https://doi.org/10.1111/j.1365-2966.2010.16713.x

[15] González-Pérez, J. N., & González-González, M. (2020). Machine and deep learning applied to galaxy morphology – A comparative study. *Astronomy and Computing, 30*, 100334. https://doi.org/10.1016/j.ascom.2019.100334

[16] Mahajan, S., et al. (2020). Galaxy morphological classification in deep-wide surveys via deep convolutional neural networks. *Monthly Notices of the Royal Astronomical Society, 491*(1), 1408–1422. https://doi.org/10.1093/mnras/stz2854

[17] Mo, H. J., Yang, X., Bosch, F. C., & Jing, Y. P. (2004). The dependence of the galaxy luminosity function on large-scale environment. *Monthly Notices of the Royal Astronomical*

Society, 349(1), 205–212. https://doi.org/10.1111/j.1365-2966.2004.07485.x

[18] Liske, J., & Brüggen, M. (2016). Galaxy classifications with deep learning. *Proceedings of the International Astronomical Union, 12*(S325), 217–220. https://doi.org/10.1017/S1743921316012771

[19] Raja, M., Hasan, P., Mahmudunnobe, M., Saifuddin, M., & Hasan, S. N. (2024). Membership determination in open clusters using the DBSCAN Clustering Algorithm. *Astronomy and Computing, 47*, 100826. https://doi.org/10.1016/j.ascom.2024.100826

[20] Xu, D., & Zhu, Y. (2024). Surveying image segmentation approaches in astronomy. *Astronomy and Computing, 48*, 100838. https://doi.org/10.1016/j.ascom.2024.100838

[21] Shamir, L. (2023). Outlier galaxy images in the Dark Energy Survey and their identification with unsupervised machine learning. *Astronomy and Computing, 43*, 100712. https://doi.org/10.1016/j.ascom.2023.100712

[22] Kramer, O., Gieseke, F., & Polsterer, K. L. (2013). Learning morphological maps of galaxies with unsupervised regression. *Expert Systems with Applications, 40*(8), 2841–2844. https://doi.org/10.1016/j.eswa.2012.12.002&#8203;:contentReference[oaicite:6]{index=6}

[23] Logan, C. H. A., & Fotopoulou, S. (2020). Unsupervised star, galaxy, QSO classification: Application of HDBSCAN. *Astronomy & Astrophysics*, 633, A154. https://doi.org/10.1051/0004-6361/201936648

[24] Jaimes-Illanes, G., Parra-Royon, M., Darriba-Pol, L., Moldón, J., Sorgho, A., Sánchez-Expósito, S., Garrido-Sánchez, J., & Verdes-Montenegro, L. (2025). Classification of HI galaxy profiles using unsupervised learning and convolutional neural networks: A comparative analysis and methodological cases of studies. *arXiv*. https://doi.org/10.48550/arXiv.2501.11657&#8203;:contentReference[oaicite:12]{index=12}

[25] Hocking, A., Geach, J. E., Sun, Y., & Davey, N. (2017). An automatic taxonomy of galaxy morphology using unsupervised machine learning. *Monthly Notices of the Royal Astronomical Society*. https://doi.org/10.48550/arXiv.1709.05834

[26] Fielding, E., Nyirenda, C. N., & Vaccari, M. (2022). The classification of optical galaxy morphology using unsupervised learning techniques. *Proceedings of the 2022 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, 1-6. https://doi.org/10.48550/arXiv.2206.06165

[27] ramacere, A., Paraficz, D., Dubath, P., Kneib, J.-P., & Courbin, F. (2016). ASTErIsM: Application of topometric clustering algorithms in automatic galaxy detection and classification. *Monthly Notices of the Royal Astronomical Society, 463*(3), 2939–2957. https://doi.org/10.1093/mnras/stw2103

[28] Ma, X., Li, X., Luo, A., Zhang, J., & Li, H. (2023). Galaxy image classification using hierarchical data learning with weighted sampling and label smoothing. *Monthly Notices of the Royal Astronomical Society, 519*(3), 4765-4779. https://doi.org/10.1093/mnras/stac3770

[29] Elvitigala, A., d. Navaratne, U., Rathnayake, S., & Dissanayaka, K. (2024). Galaxy clustering and classification using machine learning algorithms and XAI. *Proceedings of the 9th International Conference on Information Technology Research (ICITR)*, Colombo, Sri Lanka, 1-6. https://doi.org/10.1109/ICITR64794.2024.10857763

[30] Fotopoulou, S. (2024). A review of unsupervised learning in astronomy. *Astronomy and Computing, 48*, 100851. https://doi.org/10.1016/j.ascom.2024.100851