Explainable Galaxy Interaction Prediction with Hybrid Attention Mechanisms

Sathwik Narkedimilli [†], Satvik Raghav[‡], Om Mishra[*], Mohan Kumar[‖], Aswath Babu H[§], Tereza Jerabkova[**], Manish M[¶], and Sai Prashanth Mallellu[††]

[†]Department of Electrical and Computer Engineering, National University of Singapore (NUS), Singapore
[‡]School of Computing Technologies, RMIT University, Melbourne, VIC 3000, Australia
[*]Department of Electronics and Communication Engineering, Indian Institute of Information Technology Dharwad, India
[‖]Senior Data Engineer at SDI, Indian Air Force (IAF), India
[§]Department of Arts, Science, and Design, Indian Institute of Information Technology Dharwad, India
[**]Department of Theoretical Physics and Astrophysics, Faculty of Science, Masaryk University, Czech Republic
[¶] NVIDIA, India
[††]Department of Computer Science & Engineering, Symbiosis Institute of Technology, Hyderabad Campus, Symbiosis International University, Pune, India

Emails: sathwik.narkedimilli@ieee.org; satvikraghav007@gmail.com; 23bec035@iiitdwd.ac.in; mk.elitetech@gmail.com; aswath@iiitdwd.ac.in; tereza.jerabkova@eso.org; mmodani@nvidia.com; saiprashanth08@ieee.org

***Abstract*—Galaxy interaction classification remains challenging due to complex morphological patterns and the limited interpretability of deep learning models. We propose an attentive neural ensemble that combines AG-XCaps, H-SNN, and ResNet-GRU architectures, trained on the Galaxy Zoo DESI dataset and enhanced with LIME to enable explainable predictions. The model achieves Precision = 0.95, Recall = 1.00, F1 = 0.97, and Accuracy = 96%, outperforming a Random Forest baseline by significantly reducing false positives (23 vs. 70). This lightweight (0.45 MB) and scalable framework provides an interpretable and efficient solution for large-scale surveys such as Euclid and LSST, advancing data-driven studies of galaxy evolution.**

***Index Terms*—Galaxy Interactions, H-SENN, Capsule Networks, Ensemble Learning, Autoencoder-Based Feature Extraction, Explainable AI (XAI)**

I. Introduction & Related Works

Galaxies, the fundamental units of the universe, display diverse morphologies that reflect their complex formation and evolutionary histories shaped by gravitational interactions over billions of years. Galaxy interactions, encompassing mergers, tidal distortions, and minor encounters, drive structural transformations such as tidal tails, bridges, and rings, while often inducing intense star formation that exposes underlying gravitational and hydrodynamic mechanisms [2], [11]. Observed across redshifts, these processes illuminate transitions between morphological types, such as spirals evolving into ellipticals, and the influence of environmental conditions on galactic evolution [3]. Large-scale surveys like the DESI Legacy Imaging Surveys and Galaxy Zoo have cataloged millions of galaxies with detailed morphological classifications [1]. However, traditional analytical methods remain limited in their ability to predict interaction outcomes or to link them to specific physical drivers, leaving key astrophysical connections unexplored.

To address this gap, our research aims to enhance predictive accuracy while maintaining astrophysically interpretable models for galaxy interactions. With upcoming missions such as Euclid and LSST expected to produce vast datasets, the need for automated yet transparent modeling tools is increasingly critical. Although machine learning models achieve high accuracy in static morphology classification, their black-box nature limits scientific insight by obscuring causal reasoning [6]. Our study seeks to overcome this limitation by forecasting interaction outcomes, such as merger likelihood and morphological evolution, while uncovering the governing physical processes, thereby improving trust and facilitating the integration of explainable models into observational pipelines and cosmological simulations [4]. Our research introduces the following key contributions to the field of galaxy interaction prediction:

Galaxy morphology studies have advanced significantly through machine learning and cosmological analyses, offering new perspectives on galactic evolution. Cao *et al.* [6] proposed a Convolutional Vision Transformer (CvT) that combines CNNs and transformers for large-scale morphological classification, achieving superior accuracy yet limited interpretability. Haslbauer *et al.* [7] analyzed thin disk galaxy prevalence within the ΛCDM framework, finding observed thin disks more frequent than simulations predict, exposing modeling limitations in merger and feedback processes. Laishram *et al.* [8] studied galaxies at $z \sim 1.5$ using [O II] emitters, linking filamentary environments to disturbed morphologies and enhanced star formation, offering critical observational insights into interaction-driven evolution.

Deep learning and manifold learning approaches continue to enrich research on galaxy morphology. Urechiatu *et al.* [9] improved CNN-based morphology classification, achieving high accuracy across spirals, ellipticals, and irregulars, yet

lacking explainability. Semenov *et al.* [10] employed manifold learning for unsupervised feature extraction, efficiently distinguishing morphological types while improving interpretability through latent feature visualization. Despite these advances, most existing models prioritize classification performance over causal transparency or predictive modeling of galaxy interactions. These works collectively motivate our study's focus on explainable neural ensembles that merge predictive accuracy with astrophysical interpretability in galaxy interaction prediction.

The reviewed literature highlights significant progress in galaxy morphology studies, where deep learning models by Cao et al. [6] and Urechiatu [9] achieve high classification accuracy. Semenov et al. [10] enhance interpretability through manifold learning. Observational works by Haslbauer et al. [7] and Laishram et al. [8] emphasize the roles of mergers and environmental effects. Yet, deep learning approaches remain opaque, manifold methods lack predictive interaction modeling, and observational studies are not integrated with machine learning. Our research bridges these gaps by introducing an explainable, interaction-focused model using attentive neural ensembles to unify predictive power with astrophysical interpretability. The key contributions include:

- Hybrid AG-XCaps, H-SNN, and ResNet-GRU model for galaxy interaction features.
- LIME-based interpretation of key morphological drivers.
- **Astronomy Impact:** Redshift-aware morphology and merger analysis for Euclid and LSST.

II. System Model

A. Dataset

The dataset proposed by Walmsley et al. [1] contains morphology measurements for 8.67 million galaxies within the footprint of the DESI Legacy Imaging Surveys. It includes 41 columns such as unique identifiers (`dr8Id`), celestial coordinates (`RAdeg`, `DEdeg`), and object indices (`brickid`, `objid`). The remaining attributes represent predicted vote fractions for various morphological characteristics: smooth or featured (`SFSM`, `SFFDF`, `SFAF`); disk edge-on (`DEOYes`, `DEONo`); spiral arms (`SAYes`, `SANo`); bar type (`BS`, `BW`, `BNo`); bulge size (`BSD`, `BSL`, `BSM`, `BSS`, `BSNo`); roundness (`RR`, `RIB`, `RCS`); edge-on bulge shape (`EOBB`, `EOBNo`, `EOBR`); spiral winding (`SWT`, `SWM`, `SWL`); spiral arm count (`SAC1`, `SAC2`, `SAC3`, `SAC4`, `SAC4+`, `SACCT`); and handling missing values (replacing them with zeros) (`MMiD`, `MMaD`, `MM`). Metadata confirms 41 distinct morphological features recorded across more than 8.6 million galaxy entries.

The catalog was created by training deep learning models on Galaxy Zoo volunteer responses, leveraging newly collected votes for DESI-LS DR8 images and historical votes from previous Galaxy Zoo projects. These models predict the number of volunteers who would select each answer to the morphological questions, automating and standardizing the extraction of detailed galaxy features. Crucially, since these labels are derived from crowd-sourced votes, they represent probabilistic classifications that may contain human bias or disagreement, rather than absolute physical ground truths. The methodology enables extensive sky coverage and provides valuable insights into galaxy morphology by correlating automated predictions with observed features from the DESI Legacy Imaging Surveys.

B. Data Pre-Processing and Feature Engineering

Data preprocessing begins by handling missing values (replacing them with zeros), removing duplicates, converting data types, and removing outliers. The target variable `InteractionActivity` is derived by comparing the `MNo` feature to the maximum of `MMiD`, `MMaD`, and `MM`—with the result converted into a categorical format—and the dataset is further preprocessed by creating a new target variable `InteractionActivity` based on the following mathematical assumption:

$$\text{InteractionActivity} = \begin{cases} 0, & \text{if MNo} > \max\{\text{MMiD, MMaD, MM}\} \\ 1, & \text{otherwise.} \end{cases}$$

In this step, the variable `InteractionActivity` is defined by comparing the `MNo` feature against the maximum value of the features `MMiD`, `MMaD`, and `MM`. These features are described as follows: In the dataset, four columns represent the likelihood of a galaxy being involved in a galaxy interaction: MNo (Galaxy Interaction None Fraction), which is the fraction of volunteers who classified the galaxy as not undergoing any galaxy interaction; MMiD (Galaxy Interaction Minor Disturbance Fraction), representing those who identified a minor disturbance indicative of an early-stage galaxy interaction; MMaD (Galaxy Interaction Major Disturbance Fraction), representing those who classified the galaxy as undergoing a major disturbance with significant structural changes due to an ongoing interaction; and MM (Galaxy Interaction Fraction), indicating the fraction of volunteers who explicitly identified the galaxy as involved in a galaxy interaction, suggesting a late-stage interaction.

The reasoning behind the assumption is that if the probability of no interaction (MNo) exceeds the maximum likelihood among the indicators for disturbances or interactions (MMiD, MMaD, MM), the galaxy is classified as 0 (no galaxy interaction); otherwise, it is classified as 1 (galaxy interaction present). This approach simplifies galaxy categorization based on volunteer assessments and facilitates subsequent analyses using the dataset's other features.

The dataset is split into training and test sets in an 80–20 ratio, and the input variables are standardized using the StandardScaler. Feature engineering enhances data interpretability and model readiness through a structured workflow. Key predictors are first identified using Linear Discriminant Analysis (LDA) or Random Forest feature importance plots. Principal Component Analysis (PCA) is then applied to reduce dimensionality to 29 components, retaining maximum variance while minimizing redundancy. The optimized feature set is subsequently used to develop and evaluate predictive models for galaxy-interaction classification, achieving improved accuracy and computational efficiency.