



Article

Deep Learning Framework for Galaxy Morphology Classification and Analysis

Mohamed Essa ¹, Ahmad Yousef Alkahlout ¹ and Mohmmmed Taleb Mohammed ¹

¹ College of Informatics, Midocean University, Comoros; xxxxxx@st.midocean.edu.km,
253000753@st.midocean.edu.km, 252000097@st.midocean.edu.km

Abstract

Classification of galaxy morphology continues to be one of the essential topics in astrophysics, as the galaxy shape allows obtaining information about the galaxy formation process and its evolution. Due to the increasing availability of data on galaxies obtained from SDSS and DECaLS surveys, the amount of galaxy images has considerably increased. In this regard, manual visual classification becomes increasingly ineffective in the case of extensive research, highlighting the increasing relevance of automation using deep learning algorithms. Earlier works have shown that deep learning models can demonstrate good results in galaxy image classification. At the same time, many of them focus on classification accuracy, while other essential factors, such as interpretability, robustness, and prediction uncertainty, may be considered separately. For this particular research, a methodology that uses a deep learning architecture for galaxy morphology classification is analyzed. This methodology includes the usage of CNNs alongside transfer learning architectures like ResNet and EfficientNet. For the data sources, the Galaxy10 DECaLS dataset is used because of the labeled images and its widespread application in previous researches on this topic. Besides, for a more thorough representation of galaxy morphologies, Galaxy Zoo vote fraction annotations are included in this research going from single label classification to confidence aware classification of morphologies. Accuracy and F1-score are used to determine the performance of hard label classifications while RMSE and MAE are used to test the model for Galaxy Zoo vote fractions prediction problem. Moreover, the experiment tests the reliability of the model by conducting experiments based on controlled image transformations, confidence-based predictions, and visual explanations. Grad-CAM and saliency maps are used to check if the model is focused on relevant parts of galaxies. The goal of this research not only is the creation of an efficient classification algorithm but also creating a robust and interpretable methodology for dealing with astronomical images.

Keywords: galaxy morphology; deep learning; Galaxy Zoo; Galaxy Zoo Challenge; Galaxy10 DECaLS; vote-fraction prediction; confidence-aware classification; convolutional neural networks; transfer learning; explainable artificial intelligence; model robustness.

1. Introduction

Galaxy morphology classification is an important topic in astronomy because the visible structure of a galaxy can tell us a great deal about how that galaxy formed and how it changed over time. As astronomical research developed, scientists found that features such as shape, brightness distribution, spiral arms, and central bulges could provide useful clues about the physical nature and evolutionary history of galaxies. For this reason, morphology classification has remained a central task in astrophysical research [1].



Figure 1. Example of galaxy morphology complexity relevant to automated galaxy classification studies.
Source: NASA/Hubble.

As shown in Figure 1, there is significant variation within galaxy images. There are cases where galaxies show very distinguishable structures while others may be more complex or less distinct. There could be differences in the form, direction, brightness, or internal properties of the galaxies that may make classification difficult, especially when the lines between different types of galaxies are blurred.

Traditionally, the classification of galaxies was done manually using visual inspection by astronomers. Although this technique helped advance morphological study, it was labor-intensive and required a lot of time. Advances in astronomy have led to major sky surveys such as SDSS and DECaLS, greatly increasing the number of available galaxy images. As a result, manual classification has become more difficult and time-consuming [1].

As a result, there has been a lot more interest in machine learning techniques in recent times. The use of deep learning and convolutional neural networks (CNNs) shows promise for galaxy classification because they can learn features from unprocessed images. Several previous studies have reported encouraging results and have shown that CNN-based methods can perform at a level that is comparable to human classification in some tasks [2].

At the same time, there is growing interest in looking beyond classification accuracy alone. In scientific applications, it is not enough for a model to produce a correct label most of the time. They also need to understand whether the model is focusing on meaningful features of the galaxy, whether its predictions remain reliable when imaging conditions change, and whether its confidence is reasonable when the image is ambiguous. For this reason, explainability and robustness have become important parts of modern research in galaxy morphology classification [3].

The current study explores galaxy morphology classification in terms of deep learning with emphasis placed on both predictability and other aspects like robustness, uncertainty estimation, and interpretability. The primary data source for this analysis will be the Galaxy10 DECaLS galaxy morphology dataset because of its availability in labeled form to conduct supervised deep learning experiments. In doing so, it is intended to investigate the extent to which all the mentioned aspects can be utilized in creating an effective solution for automated galaxy morphology classification [3].

Various works have studied several ways of classifying galaxies through deep learning, such as semi-supervised learning, augmentation-based solutions, and transformer models. All the above areas are pertinent to the current study to a certain extent. It should be noted that the current study does not focus on each area as a separate research problem but uses a selected portion of modern architectures and preprocessing techniques depending on their ability to contribute to the main goal of developing a viable framework for galaxy morphology [8].

In particular, in such a context, the value of deep learning model application in astronomical studies cannot be reduced only to high prediction capability. The use of such models can bring additional benefits like a better understanding of the process behind their decisions, stability in different image settings, and, finally, the representation of the uncertainty associated with galaxy structure classification. Therefore, galaxy morphology classification in the study will be addressed from the angle of its performance, reliability, and interpretability [13].

1.1. Problem Statement

Many deep learning studies have applied deep learning for galaxy morphology classification, yet the majority of them pay attention only to accuracy and F1-score measures. Though these measures provide certain information, they do not cover all the aspects important for scientific analysis.

One of these aspects is interpretability. For example, it would be interesting to see how deep learning model interprets galaxy morphology. When classifying the galaxy as spiral, one wants to check that this result is based on spiral structure and central bulge rather than any other feature present in the image.

The second issue refers to variability of galaxy images observed with different conditions. Differences in illumination, noise, contrasts and so on can influence the picture significantly. Therefore, the classification performed by the model will demonstrate poor results in case the image is altered in any way. Scientific applications require robust and reliable prediction even for changed images.

Uncertainty may also become a critical factor for galaxy morphology classification since some galaxies may lack morphological features while the others can contain features corresponding to several classes. As a result, labeling only by final class may be not enough. The more helpful information the model provides about the uncertainty degree, the more confident in its predictions the scientist will be.

Another problem associated with this issue is that using only one final label for every galaxy does not properly account for the uncertainty of some morphologies. There are galaxies having combined morphologies and, in such cases, confidence-based representations provide more information than simply

using labels. This calls for designing a method which can start from class-label classification but at the same time allows richer representations in case they are available.

Based on the above, the main problem considered in this study can be formulated as follows: current deep learning solutions for galaxy morphology classification provide accurate predictions but fail to provide a unified framework for classification, which includes interpretability, robustness, and prediction uncertainty.

It means that there is a significant difference between achieving high classification accuracy and creating a reliable classification model, which explains its decisions and prediction uncertainty degree. Therefore, a more complex solution should be suggested that will satisfy all three criteria at once.

For testing the solution, the Galaxy10 DECaLS dataset is used as a primary source of labeled images suitable for deep learning experiments.

1.2. Research Objectives

The goals of the present study can be formulated as follows:

- Create baseline CNNs models to classify galaxy morphology and assess their capability of learning the features directly from the images.
- Use transfer learning approach and pretrained models (such as ResNet and EfficientNet) to improve the models' performance and generalization abilities especially in case of lack of training samples.
- Develop methods of galaxy morphology classification from controlled hard-labeling approach to confidence-based approach with Galaxy Zoo vote-fraction annotations where applicable.
- Evaluate classification performance using quantitative measures (accuracy and F1-score) in order to test the ability of the models to classify galaxy types.
- Test model robustness with respect to controlled perturbations introduced into images and assesses the effect of different observational conditions on the models' predictions.
- Consider the confidence and uncertainties involved in predicting the form of galaxies to increase our knowledge about the reliability of the model when dealing with uncertain forms of galaxies.
- Apply an Explainable AI approach, like Grad-CAM and saliency maps, to check whether the model is focusing on scientifically relevant structures of galaxies while making predictions.

1.3. Research Contribution

This research offers several key contributions to the study of automated galaxy morphology classification:

- A structured evaluation of deep learning architectures:

The work provides a systematic assessment of different deep learning models applied to galaxy morphology classification, including both baseline CNNs and transfer-learning models. This helps clarify how various architectures perform on astronomical image data.

- A confidence-aware extension of the classification framework:

The research expands the scope of the method from a class-labeled approach to galaxy morphology classification to the use of the vote fraction representation. Such an approach allows investigating galaxy morphology with uncertain characteristics and combining classical classification with the more sophisticated hierarchy-based description of galaxy structure.

- A robustness evaluation procedure:

The study introduces a method for examining how model performance changes under controlled variations in the input images. This robustness analysis highlights the sensitivity of the models to observational differences and helps determine how stable their predictions are.

- Estimation of uncertainty of the model prediction:

The estimation of uncertainty is included in the research to reveal how confident the model is at predicting certain ambiguous or transitional galaxy morphologies.

- The usage of the explainable AI approach:

Grad-CAM is used in this case to analyze if the model focuses on the important parts of the galaxies during the classification process.

The rest of this paper is organized as follows. Section 2 presents the related work on galaxy morphology classification using deep learning approaches. Section 3 describes the proposed methodology, including datasets, preprocessing, model families, and evaluation procedures. Section 4 presents the experimental results and discussion. Finally, Section 5 concludes the paper and outlines future research directions.

2. Related Work

The following section gives a brief introduction of the background information and previous works about galaxy morphology classification using deep learning approaches. First, this section introduces some basic terminologies in terms of galaxy morphology classification along with its history from visual inspection to the application of automatic methods. Then, various existing studies that use deep learning to process galaxy images are presented briefly.

As mentioned in Section 1, there is an increased number of works done in recent years regarding galaxy morphology classification thanks to the availability of large-scale image databases due to astronomical surveys. In parallel with the increasing number of works, deep learning models, specifically convolutional neural networks, have proven their ability to capture meaningful features of images, thus they become a promising choice for galaxy morphology classification.

However, at the same time, most existing works concentrate on accuracy of galaxy morphologies classification. Although accuracy is one of the crucial criteria in evaluating any classification model, in scientific domains, other aspects including generalization capabilities, uncertainty quantification, and explainability are also very important in applying deep learning algorithms.

Therefore, in addition to focusing on performance measures, this section will discuss previous approaches in detail in terms of methodology, datasets used, and evaluation metrics in order to identify their advantages and disadvantages as well as potential research gaps for future investigations.

The reviewed literature indicates that there have been great strides made in the field of galaxy morphological classification. Nonetheless, it is apparent that majority of these prior works have mainly focused on one aspect of the problem. There are those who prioritize accuracy, those who consider the importance of interpretability, and even those who talk about uncertainties and limitations in the data.

Walmsley et al. [1] Walmsley investigated the morphological classification of galaxies on a larger scale using machine learning techniques on current astronomical datasets. The use of the Galaxy Zoo DECaLS dataset allowed the researchers to utilize morphological classifications provided by many citizen scientists. The authors developed a deep learning model called Zoobot, which integrates convolutional neural network and probability approaches to estimate the entire distribution of answers provided by volunteers for about 300,000 galaxies. The model provided good agreement with human classification efforts and

demonstrated good prediction accuracy, even regarding complicated morphological patterns. The main advantage of the investigation under analysis is the creation of one of the biggest morphological catalogs created using deep learning techniques. However, the use of volunteer-generated answers can be considered the main disadvantage of the methodology employed in this research.

The significance of this line of investigation to the current research is that it views galaxy morphology as a probabilistic problem in terms of distributions rather than just a hard-labeling issue. The reason why this perspective is important is that the current study begins with a controlled hard-labeling issue through the Galaxy10 DECaLS dataset but also involves the bigger Galaxy Zoo vote fraction problem as well. In this regard, previous work done on Galaxy Zoo provides a useful foundation for expanding galaxy classifications beyond one label.

Dieleman et al. [2] In their paper, Dieleman et al. proposed a deep learning framework for galaxy morphological classification using rotation-invariant CNNs. The study applied images from the Galaxy Zoo data set and was aimed at developing new CNN architecture which would allow classifying galaxies irrespective of their rotational orientation. With the introduction of rotation symmetry to the architecture of the network, it was able to better learn the morphological patterns, including spiral arms and bulge characteristics of galaxies, directly from the pixel-level data. According to the authors, the proposed method reached human-like accuracy levels in morphological classification, achieving about 91% of prediction accuracy. An obvious advantage of this approach is in its ability to handle rotational variability, characteristic of many astronomical images. However, as compared to standard CNN models, additional complexity arises due to the introduction of the new property.

Bhambra et al. [3] conducted research on the explainability of deep learning models which were used in the classification of galaxies based on their morphology. In this research, different explainable artificial intelligence approaches such as Grad-CAM and saliency maps were used to illustrate image areas that contributed towards the prediction made by deep learning models. Through galaxy images obtained from large sky surveys such as Galaxy Zoo and DECaLS, it was shown that deep neural networks were able to recognize different morphological structures within galaxies such as spiral arms, galactic core structures, and even asymmetry. This research illustrates that using explainability techniques, it is possible to learn more about the internal functioning of deep learning models.

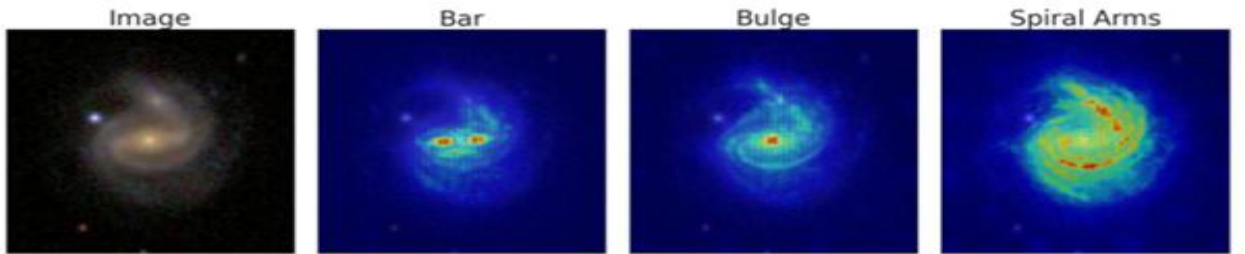


Figure 2. Example of a saliency map highlighting discriminative regions in galaxy morphology classification. Source: Adapted from Bhambra et al. (2022) [3].

As shown in Figure 2, saliency maps highlight the regions of the galaxy image that contribute most strongly to the model's prediction. In many cases, the visualization confirms that the neural network focuses on key morphological structures such as spiral arms, bars, or the central bulge. This provides an additional layer of reliability by allowing astronomers to verify whether the classification decision is based on meaningful physical features. In this study, a similar explainability approach will be used to evaluate

whether the proposed deep learning framework focuses on relevant morphological structures when classifying galaxy images.

Domínguez Sánchez et al. [7] researched the application of deep convolutional neural networks in improving the accuracy of galaxy morphology classification using images captured by the Sloan Digital Sky Survey (SDSS). The model they created aimed at being able to recognize the important attributes used to differentiate between certain types of galaxies like elliptical and spiral galaxies. With the aid of the large number of images provided by the SDSS survey, the CNN model managed to classify the images with a success rate of about 95%. One strength of the work by Dominizes Sanches is that the CNN does not depend on a predefined set of features, as it can learn important morphological characteristics directly from the pixel data.

Luo et al. [8] believed problems related to galaxy morphological classification in situations when there are not many labeled samples available a frequent situation in the field of astrophysics. They developed an innovative semi-supervised deep learning algorithm that uses convolutional neural network and GAN models. In this case, the experiments were conducted using the Galaxy10 DECaLS dataset, and the classification accuracy obtained was around 84%. One positive point about this study is the reduced need for big fully labeled samples due to the simultaneous application of both labeled and unlabeled data. Such an innovation will be particularly beneficial in astronomy due to the time consumption involved in the process of labeling. On the other hand, one drawback of such an algorithm is its instability.

Yao et al. [9] focused on improving galaxy morphology classification through data augmentation when there is limited data for training. In this study, the Galaxy10 DECaLS dataset was used, in which a few-shot-learning-inspired augmentation technique was applied to enhance image diversity within the training set. This research indicated that with their augmentation technique, the robustness of their models could be increased to achieve up to 95% accuracy rate. One major advantage of this research is that they managed to show that augmentation can be effective without requiring changes to the core classification problem. On the other hand, one major limitation of this study is that overuse of augmentation or unrealistic augmentation might lead to spurious patterns.

Le et al. [10] carried out a comparative study of several deep learning models for galaxy morphology classification. The researchers employed the Galaxy10 DECaLS dataset to test customized CNN algorithms against the transfer learning algorithm models like ResNet and EfficientNet. It is demonstrated that the EfficientNet model produced an accuracy score of about 87%, which is considered among the best-performing models tested in the paper. The main advantage of this study lies in the fact that it presents a comparative analysis between several architectures under identical experimental conditions. In addition, experimental research was limited to utilization of just one dataset, implying its potential applicability to other astronomical survey data.

Xie et al. [11] conducted research on the challenge of classifying galaxy morphology with the use of a deep convolutional neural network inspired by the VGG19 architecture. For experimentation, the researchers applied the Galaxy10 DECaLS dataset, which contains labeled galaxy images taken through extensive astrophysical observation processes. Leveraging transfer learning techniques, the researcher were able to train the VGG19 model to recognize hierarchical visual features to distinguish between the different types of galaxies. As a result, the proposed approach managed to achieve about 84.5% accuracy, showing that the application of deep CNNs is adequate for this type of problem. At the same time, the authors identified that models like VGG19 have a relatively high number of parameters, which might be responsible for overfitting.

Cheng [12] explored the effectiveness of using Vision Transformers (ViT) to classify galaxy morphology. Specifically, the Galaxy10 DECaLS dataset was used to investigate whether transformer models outperform regular CNNs in terms of their ability to find global features within images. By leveraging the self-attention mechanism, the model in question highlighted key areas of the galaxy to increase overall accuracy up to 86%. An interesting outcome of this research is that transformers can detect dependencies throughout the whole image rather than within specific parts only. A downside observed by Cheng was the significant amount of training data needed.

Cao et al. [13] came up with a combination approach to deep learning where the researcher merged CNNs with elements from transformers to classify the morphology of galaxies. For this research, Cao used galaxy images from the Galaxy Zoo database. Specifically, CvT architecture was used because the model integrates local feature detection in CNNs and global feature detection in transformer architectures. The achieved accuracy was around 98%, thus making it one of the most effective models among those that have been reviewed in this analysis. One strength of this research paper is that it uses both architectural designs, hence can detect patterns in galaxies using a wide range of approaches. However, the improvement made in the model makes it complex, thus requiring more computational resources.

Urechiatu et al. [14] focused on the potential usefulness of convolutional neural networks (CNNs) in improving automated classifications of galaxy morphology. For the experiment, the authors used a dataset consisting of 10,000 images of galaxies taken from the Galaxy Zoo 2 database, which is known to contain labeled images provided by many volunteers. Specifically, to build the algorithm, the authors came up with a CNN architecture which they expected to be able to distinguish morphological traits better and applied a number of methods to preprocess and augment data. As a result, the accuracy in classification reached 92%. However, despite high effectiveness in recognizing simple structures of galaxies, the models failed to work properly when dealing with complicated morphologies.

Bom et al. [15] explored analyzing deep learning algorithms' capabilities for performing morphological classification of galactic objects with the help of imaging data obtained during the S-PLUS sky survey. However, apart from relying on the data set described above, the researchers suggest including other sources of information in the analysis, namely morphological catalogs resulting from other big astronomical projects. According to the study, CNNs that can detect morphological features in input images could be useful tools to accomplish the proposed task. Namely, it means that a neural network should be capable of detecting certain morphological characteristics of galaxies' core, light distribution in galaxies, as well as possible spirals. As a result, the network can recognize different types of galaxies accurately enough, with the accuracy of classification being 92% to 95%. Therefore, CNNs can be considered effective tools for accomplishing the described tasks, given proper training of the network. Still, certain observational conditions may impact CNNs' work negatively.

Wei et al. [16] used transformer architecture to analyze galaxy morphology classification. The experiment was done on the Galaxy10 DECaLS dataset, which consists of galaxy images labeled after extensive surveys done by astronomers. This experiment used (ViT), where self-attention is used to recognize global patterns within the galaxy images. The results obtained show an accuracy rate of about 94%, thus the applicability of the transformer architecture to such tasks. One major strength of using this architecture to classify galaxy morphologies is that it can learn long-range dependencies in data. On the other hand, transformers need huge datasets to perform at their best.

Almeida et al. [17] examined the role of transfer learning in enhancing galaxy morphology classification using deep convolutional neural networks. This research utilized data from the Galaxy Zoo project, which

contains images labeled using massive citizen scientist inputs for galaxy morphology. Using the transfer learning method, fine-tuning the pre-trained ResNet model allowed the neural network to learn to recognize the differences between galaxy morphologies, such as spirals and ellipticals. Based on the study findings, the transfer learning method used yielded an average of about 92% accuracy, showing the usefulness of transfer learning with large datasets of natural images to improve results in astrophysics. However, one challenge in performing such transfer learning is the class imbalance in the dataset since morphological differences will be harder to learn by the neural network.

Zhang et al. [18] explored the utilization of a new machine learning approach to classify galaxy morphology. This novel technique was a combination of the convolutional neural network and attention mechanisms. The Galaxy10 DECaLS dataset was used, containing labeled galaxy images obtained from various astronomical researches. The use of the attention mechanism made it possible for the model to focus more on the key features in the images. The accuracy achieved in the study was around 96%, which shows that the technique might improve classification effectiveness.

It is seen from the analyzed literature that DL is a promising method for classifying galaxies according to morphological types. The above-mentioned literature employs various techniques, including classical CNN models, transfer learning, semi-supervised learning, attention-based methods, and transformer architecture among others. In general, the considered literature proves that application of DL to galaxy morphology classification yields promising results provided that a large set of astronomical images is available.

However, even with the common high performance of deep learning techniques, each technique has its advantages and drawbacks. For example, some papers are oriented towards improving the accuracy, whereas others pay more attention to other aspects of galaxy morphology classification, such as interpretability, estimation of uncertainties, or the influence of training data insufficiency. In most cases, these features were investigated separately without creating an integrative approach.

The comparison of the selected studies is presented in Table 1.

Table 1. Comparison of related studies in galaxy morphology classification

Ref	Author	Year	Technique Used	Accuracy	Dataset	Comparison with This Research
[1]	Walmsley et al.	2022	Bayesian Deep Learning (Zoobot)	Not explicitly reported	Galaxy Zoo / DECaLS	Reference for uncertainty-aware galaxy classification
[2]	Dieleman et al.	2015	Rotation-Invariant CNN	~91%	Galaxy Zoo	Baseline CNN architecture for morphology prediction
[3]	Bhambra et al.	2022	Grad-CAM & Saliency Mapping	Not focus of study	Galaxy Zoo / DECaLS	Reference study for explainable AI evaluation
[7]	Domínguez Sánchez et al.	2018	Deep Convolutional Neural Network (CNN)	~95%	SDSS Galaxy Dataset	Deep learning benchmark for galaxy morphology classification

[8]	Luo et al.	2025	Semi-supervised Learning (GAN-based)	~84%	Galaxy10 DECaLS	Semi-supervised learning reference under limited labels
[9]	Yao et al.	2024	DCGAN Data Augmentation	Up to ~95%	Galaxy10 DECaLS	Data augmentation reference for improving model generalization
[10]	Le et al.	2025	EfficientNetB5 Transfer Learning	~87%	Galaxy10 DECaLS	Transfer learning architecture benchmark
[11]	Xie et al.	2024	VGG19 Transfer Learning	~84.5%	Galaxy10 DECaLS	CNN transfer learning baseline comparison
[12]	Cheng	2022	Vision Transformer	~86%	Galaxy10 DECaLS	Transformer-based architecture comparison
[13]	Cao et al.	2024	Convolutional Vision Transformer (CvT)	~98%	Galaxy Zoo CANDELS	Hybrid CNN–Transformer performance benchmark
[14]	Urechiatu et al.	2024	Convolutional Neural Network (CNN)	~92%	Galaxy Zoo 2	CNN-based classification improvement reference
[15]	Bom et al.	2024	Deep Neural Network (CNN-based)	~92–95%	S-PLUS astronomical survey	Baseline CNN architecture for morphology prediction
[16]	Wei et al.	2024	Vision Transformer (ViT)	~94%	Galaxy10 DECaLS	Transformer-based architecture comparison
[17]	Almeida et al.	2024	Transfer Learning (ResNet CNN)	~92%	Galaxy Zoo	Deep learning benchmark for galaxy morphology classification
[18]	Zhang et al.	2024	CNN with Attention Mechanism	~96%	Galaxy10 DECaLS	CNN with attention mechanism benchmark

Although recent studies have reported promising results in galaxy morphology classification, several important challenges remain. Many existing models achieve good accuracy, but they do not always provide enough information about how stable their predictions are when image conditions change. In addition, interpretability and uncertainty estimation are still not fully included in many classification frameworks.

Another issue is that some studies depend heavily on large, labelled datasets, while future astronomical surveys may continue to produce data at a scale that makes full labelling difficult. Differences in image quality, brightness, noise, and observational conditions can also affect model performance and reliability.

For these reasons, there is still a need for a framework that does more than classify galaxy images accurately. A more useful approach should also consider robustness, uncertainty, and interpretability

together. This is the main gap that motivates the current research, which aims to study galaxy morphology classification through a framework that combines these important aspects in one setting.

Besides those constraints already mentioned, there is another issue with many of the studies in the current literature, which regard classification accuracy, uncertainty, robustness, and interpretability as distinct paths of research. One of the other challenges is that of the gap between the hard-label classification approach in galaxy morphology data and the richer one of vote fraction representation. This is why the current study attempts to bridge the gap between class label classification and a more general confidence-based approach.

3. Proposed Methodology

This section describes the methodological approach used to create the proposed galaxy morphology classification and analysis system. Specifically, this section covers the key phases of the research process including data input, preprocessing, model creation, training approach, evaluation process, and reliability analyses. The aim is to show how the experimental design and its controls were established prior to discussing the results in the next section.

The proposed methodology is designed as a sequential pipeline. First, input data are processed and split into relevant experimental subsets. Second, preprocessing is conducted to make the images and labels suitable for deep learning models. Third, a few models are developed based on the differences in their methodological role. Finally, models are tested using the appropriate performance and reliability metrics.

Figure 3 provides a schematic illustration of the overall methodological pipeline. The details of the datasets used in this experiment are provided separately in Section 3.1.

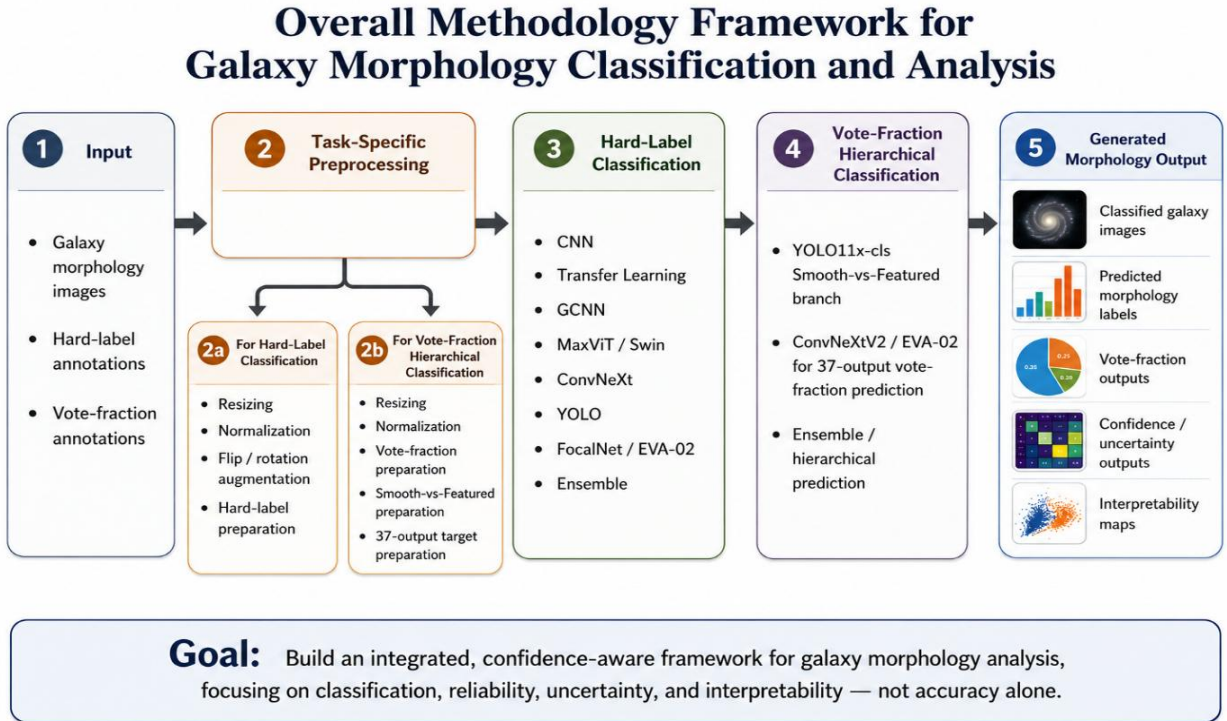


Figure 3. Overall proposed galaxy morphology methodology framework.

3.1. Input (Datasets)

Two interrelated datasets were used in this study. Galaxy10 DECaLS was used as the controlled dataset for hard-label ten-class galaxy morphology classification. Galaxy Zoo The Galaxy Challenge was used as a larger galaxy zoo dataset for vote-fraction-based hierarchical prediction. Both datasets play distinct methodological roles in the context of the same galaxy morphology trajectory.

3.1.1. Galaxy10 DECaLS Dataset

Galaxy10 DECaLS is a supervised galaxy image dataset from the Dark Energy Camera Legacy Survey. The dataset contains 17,736 galaxy images belonging to ten morphology classes. Each image has an original size of $256 \times 256 \times 3$ pixels. This dataset serves as the foundational hard-label classification dataset, helps in conducting evaluation of different model types, and assists in studying ambiguity patterns between morphologically related classes.

For the key Galaxy10 DECaLS experiments, an approximate 70% / 15% / 15% train/validation/test split was used. The training subset was used for parameter learning, the validation subset was used for model selection and controlled refinement, and the test subset was used for the final evaluation.

Table 2. Galaxy10 DECaLS split used in the main classification experiments.

Subset	Images	Approx. ratio	Role
Training	12,415	70%	Parameter learning
Validation	2,660	15%	Model selection and refinement decisions
Testing	2,661	15%	Held-out final evaluation
Total	17,736	100%	Full labeled Galaxy10 DECaLS dataset

Figure 4 provides examples of images from the Galaxy10 DECaLS dataset illustrating the variety of hard-label galaxy morphology classification task analyzed in this study.



Figure 4. Sample galaxy images from the Galaxy10 DECaLS dataset used for hard-label morphology classification.

3.1.2. Galaxy Zoo Challenge Dataset

Galaxy Zoo – The Galaxy Challenge allowed expanding the methodology from the hard-label classification task to the vote-fraction-based hierarchical prediction. The labeled part of the dataset contains 61,578 galaxy images annotated in the training_solutions_rev1.csv file. Each labeled image has 37 targets which correspond to the Galaxy Zoo hierarchy.

The labeled part was further divided into training, validation, and internal test subsets using the approximate 70% / 15% / 15% split. Official Kaggle test images cannot be used for evaluation because their ground truth labels are not provided. These images were not used for training, model selection, hyperparameter tuning, and internal test evaluation; they were used only for final prediction generation.

Table 3. Galaxy Zoo Challenge labeled split and official unlabeled test set role.

Galaxy Zoo subset	Images	Approx. ratio	Role
Training	43,104	70%	Parameter learning for 37-output prediction
Validation	9,237	15%	Model selection and ensemble design
Internal testing	9,237	15%	Held-out internal evaluation
Labeled total	61,578	100%	Images with vote-fraction annotations
Official Kaggle test	79,975	Not part of labeled split	Unlabeled images used only for final submission prediction

Figure 5 provides examples of images from the Galaxy Zoo Challenge dataset illustrating the vote-fraction nature of the dataset, whereby each galaxy image has probabilistic labels rather than a single label.

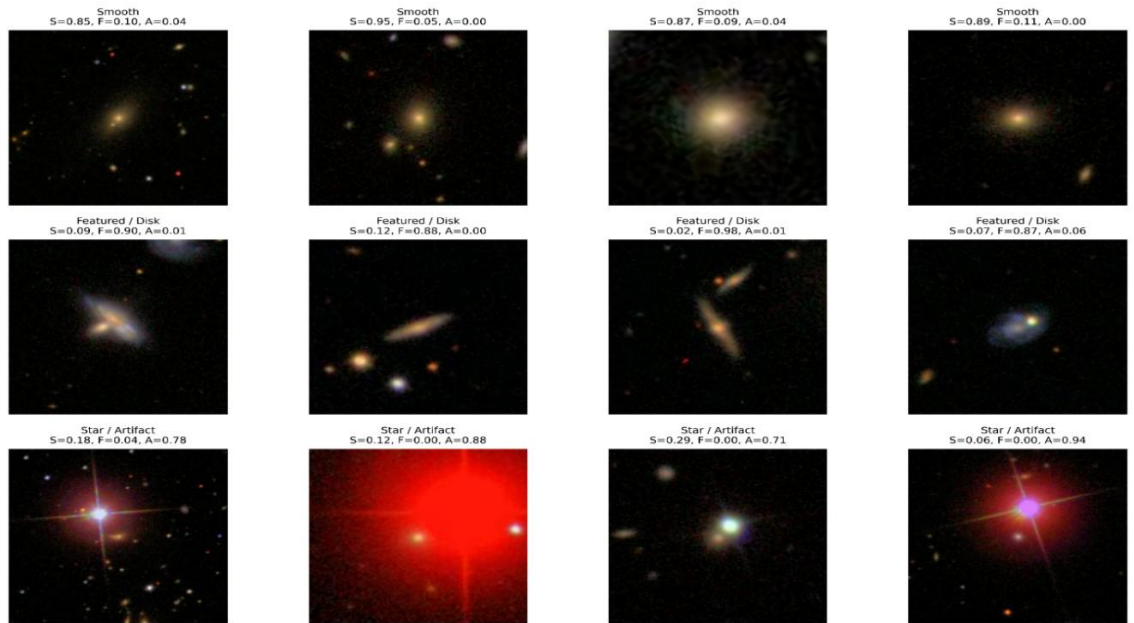


Figure 5. Sample Galaxy Zoo Challenge images with first-gate vote-fraction annotations used for hierarchical morphology classification.

3.2. Preprocessing

Preprocessing was conducted in accordance with the learning task and the structure of the labels for each dataset. Since the methodology includes both hard-label classification and vote-fraction hierarchical classification tasks, preprocessing is conducted in two main branches: galaxy images are prepared for hard-label classification and images, and vote-fraction annotations are prepared for hierarchical prediction.

3.2.1. Preprocessing for Hard-Label Classification

For the Galaxy10 DECaLS branch, the preprocessing focuses on preparing the images for supervised hard-label classification. The original images were resized to satisfy the input requirements for each model type. Different input resolutions were used during development: 64×64 , 128×128 , 160×160 , 224×224 , and 256×224 . Smaller resolutions were used during early baseline experiments; larger resolutions preserve more detail in galaxy images.

For the pretrained models, ImageNet-style normalization was applied as necessary using:

$$x_{\text{norm}} = (x - \mu) / \sigma \quad (1)$$

Where μ and σ denote the channel-wise mean and standard deviation used for normalization.

For ImageNet-pretrained models, images were normalized using standard ImageNet statistics: $\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$. In earlier baseline experiments, other normalization options were used in accordance with the model configuration. Training-time augmentation was applied to the training subset only and included transformations such as horizontal flips, vertical flips, and rotation-based augmentation. Validation and test images were not augmented in order to avoid data leakage and maintain the integrity of model selection and evaluation process.

3.2.2. Preprocessing for Vote-Fraction Hierarchical Classification

For the Galaxy Zoo Challenge branch, the preprocessing takes into account the vote-fraction structure of the labels. Each image corresponds to the set of 37 vote-fractions which correspond to the morphology questions in Galaxy Zoo hierarchy. Therefore, preprocessing involves creating a continuous target vector for each galaxy image instead of a single class label.

For the Galaxy Zoo branch, preprocessing also includes preparation of the vote-fraction targets. The first-level vote fractions were used to create the Smooth vs Featured high-confidence stage, while the full set of 37 vote-fraction targets was kept for hierarchical morphology prediction.

$$y \in [0, 1]^{37}, y = [y_1, y_2, y_3, \dots, y_{37}] \quad (2)$$

In addition to the full vote-fraction prediction task, the high-confidence Smooth vs Featured stage was created using the first-level Galaxy Zoo vote fractions. The sample was kept for this binary hierarchical task if one of the selected vote fractions satisfies the confidence threshold τ :

$$\text{label} = \text{Smooth} \text{ if } \text{Class1.1} \geq \tau; \text{ Featured if } \text{Class1.2} \geq \tau \quad (3)$$

For the high-confidence Smooth vs Featured branch, the actual implementation also included YOLO11x-cl training at 448×448 resolution, while the full Galaxy Zoo branch preserves the 37 vote-fraction targets for hierarchical morphology prediction.

This preprocessing approach allows accommodating both controlled hard-label classification and broader confidence-aware hierarchical morphology prediction.

3.3. Experimental Control and Data Partitioning

Experimental control was maintained by clearly distinguishing among the roles of training, validation, and test data. The training subset is used solely for parameter learning, the validation subset is used for model selection and controlled refinement, and the held-out test subset is used for the final evaluation. This division minimizes the chance of data leakage and ensures that the reported results are uniformly interpretable. This approach is consistent with recent studies of galaxy morphology classification which emphasize dataset splitting and held-out evaluation [8].

For the Galaxy Zoo Challenge branch, the official Kaggle test images were separated from the labeled training subset because of the lack of ground truth vote-fraction labels. The images are not used for training, model selection, internal loss calculation, or hyperparameter tuning; they are used only for the final prediction generation after the development of the models is completed.

Diagnostic branches such as reduced-class analysis and disagreement-based refinement are used as the methodological instruments for the analysis of ambiguity and errors. These branches do not re-define the final task or the evaluation process. When the prediction outputs are required to fit the external submission format, the quality control is conducted: Galaxy ID consistency, expected column number, row count, missing values, prediction format.

3.4. Classification Models

The classification models in this study were organized into several families to support the proposed galaxy morphology framework. This organization reflects the need to evaluate galaxy morphology not only as a hard-label classification problem, but also as a confidence-aware and vote-fraction-based prediction task [1].

CNN-based models were included because they are widely used for learning local visual patterns in galaxy images [2]. Transfer-learning CNN backbones were considered because previous studies have shown their usefulness for galaxy morphology classification when pretrained visual representations are adapted to astronomical images [10]. Transformer and hybrid attention models were included to examine broader spatial relationships in galaxy morphology images [12].

The purpose of using multiple model families is to compare different models and to explore how different architectural principles affect galaxy morphology recognition. Conventional CNNs serve as a controlled baseline to learn the local visual patterns from galaxy images. Transfer-learning CNN backbones use pretrained visual representation for the extraction of the image features. Transformers and hybrid attention models learn broad spatial relationships between pixels. YOLO classification models provide efficient image-level classifier, especially in the case of high-confidence morphology classification branch. Finally, ensembles and hierarchical models integrate individual model outputs with the Galaxy Zoo vote-fraction prediction task.

For an input galaxy image x , a general hard-label classification branch can be expressed as follows:

$$h = f_{\theta}(x) \quad s = W h + b \quad \hat{y} = \text{softmax}(s) \quad (4)$$

The softmax probability for class c is defined as:

$$\hat{y}_c = \frac{\exp(s_c)}{\sum_{j=1}^C \exp(s_j)} \quad (5)$$

where $f\theta(x)$ denotes the feature extractor, h is the learned feature representation, s is the logit vector, W and b are the parameters of the classification head, C is the number of classes, and $\hat{y}_c$ is the predicted probability for class c .

The above-described softmax-based formulation applies to the hard-label classification branches. For the Galaxy Zoo vote-fraction branch, the output is considered as a continuous multi-output morphology representation.

The following subsections describe each of the model families in terms of their methodological role, architectural principle, mathematical representation, and representative architectural diagrams.

Since the datasets include publicly available astronomical image collections without any human or personal data, the key methodological concern in this study is the experimental validity rather than privacy preservation. The data manipulation was conducted in order to ensure correct dataset splitting and prevent data leakage between the training, validation, and test subsets. The role of the official unlabeled Kaggle test images is also specified clearly.

3.4.1. Deep Learning CNN Family

The deep learning CNN family serves as the controlled baseline layer of the proposed galaxy morphology classification framework. CNN models are appropriate for this task because the galaxy morphology depends greatly on the local visual structures (such as spiral arms, central bulges, bars, edge-on disks, texture-like structures) which can be learned directly from galaxy images using convolutional filters. This rationale is consistent with previous work on galaxy morphology classification [2].

Within this family, the baseline CNN serves as the initial supervised architecture for the learning of the local morphology features. The model accepts preprocessed galaxy image and feeds it into a sequence of convolutional blocks. Each block applies convolutional filtering, normalization, nonlinear activation, and pooling operations if necessary. The resulting feature maps are aggregated by global average pooling and fed to the fully connected classification layer.

The convolution operation used in the CNN branch is described as follows:

$$F(i, j, k) = \sum_m \sum_n \sum_c X(i + m, j + n, c) W(m, n, c, k) + b_k \quad (6)$$

where X denotes input image or feature map, W is the convolution kernel, b_k is the bias term for output channel k , and $F(i, j, k)$ is the extracted feature value at spatial position (i, j) and channel k .

In addition to the baseline CNN, the GCNN D16 is included as the symmetry-aware extension of the same convolutional family. Galaxy images can appear in various orientations but still belong to the same morphological class. Therefore, the model capable of capturing the rotation- and reflection-related patterns is helpful in the galaxy morphology classification task. This approach is consistent with previous work on rotation-aware galaxy morphology classification, in which symmetry information helped to capture morphology features under rotation [2].

A simple form of group convolution is given by:

$$[f * \psi](g) = \sum_{x \in \mathbb{Z}^2} f(x) \psi(g^{-1}x) \quad (7)$$

where f is the input feature map, ψ is the learnable filter, and g denotes the transformation element from the symmetry group. Using this operation, the GCNN branch produces symmetry-aware feature maps before pooling and classification.

Figure 6 demonstrates the deep learning CNN family used in this study. The baseline CNN branch extracts local morphology features using conventional convolutional blocks, while the GCNN D16 branch adds symmetry-aware representation of rotated and reflected galaxy images.

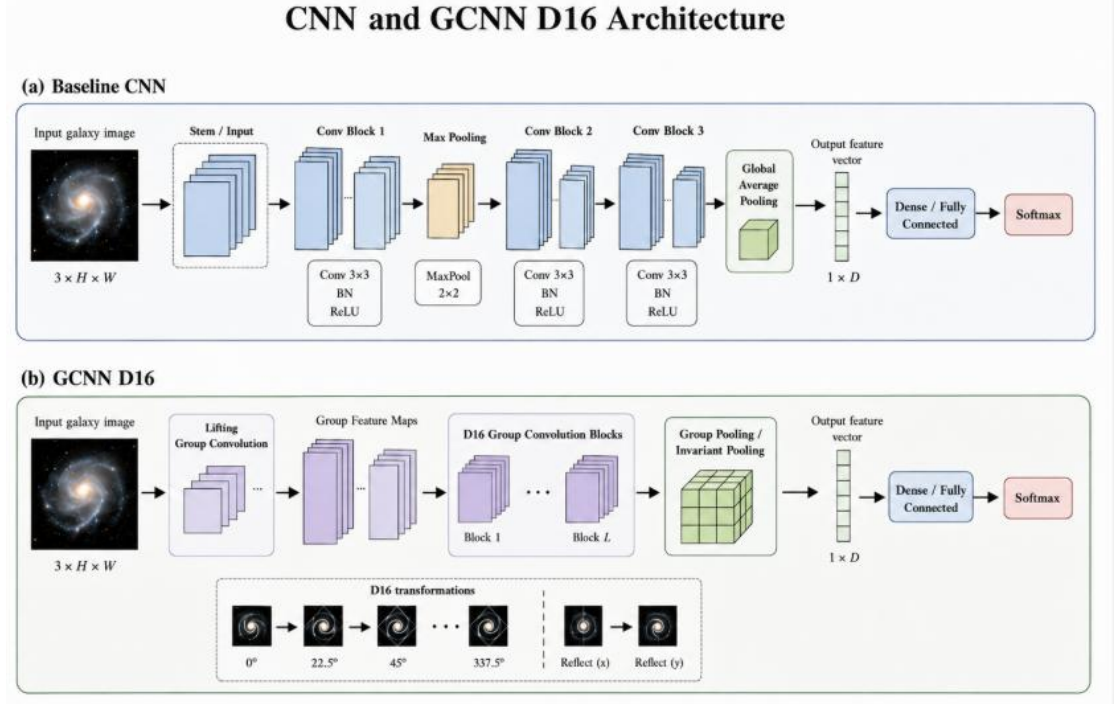


Figure 6. CNN family architecture used for local galaxy morphology feature extraction.

3.4.2. Transfer-Learning CNN Backbones Family

The transfer-learning CNN backbones family was included in order to have more robust pretrained visual representations than the baseline CNN branch. Instead of learning all visual features from the galaxy dataset alone, transfer-learning models learn the pretrained weights and adapt them to galaxy morphology classification task. Such approach is consistent with previous studies on galaxy morphology, in particular, comparison of pretrained CNN architectures and transfer-learning approaches such as EfficientNet, VGG-based models, and ResNet-based models [10].

In the proposed framework, the pretrained backbone acts as the feature extractor. The original classification layer of the pretrained model is replaced with the new task-specific classification head. The backbone extracts the compact representation of the input galaxy image, and the head maps this representation to the desired morphology output.

The general transfer-learning process can be expressed as:

$$h = f_{\theta}(x)$$

$$\hat{y} = \text{softmax}(W_{\text{new}}h + b_{\text{new}}) \quad (8)$$

where $f_{\theta}(x)$ denotes the pretrained CNN backbone, h is the extracted feature representation, and W_{new} and b_{new} are the parameters of the newly added classification head.

ResNet-style backbones were included in order to use the residual learning which facilitates training of the deep networks. A residual block is described as:

$$y = F(x) + x \quad (9)$$

where x is the input to the block, $F(x)$ is the learned residual transformation, and y is the output of the block.

EfficientNet-style backbones were included due to the compound scaling which allows balancing the network depth, width, and input resolution:

$$d = \alpha^\phi, w = \beta^\phi, r = \gamma^\phi \quad (10)$$

subject to: $\alpha \cdot \beta^2 \cdot \gamma^2 \approx 2$

where d , w , and r represent depth, width, and resolution, respectively, and ϕ is responsible for the overall scaling level.

In the transfer-learning CNN backbones family, the various CNN backbones are considered in order to provide the different ways of the feature extraction. ResNet-style models use residual connections in order to facilitate the training of the deep network, while EfficientNet-style models use the compound scaling. ConvNeXt and ConvNeXtV2 were included as the contemporary CNN backbones with integration of convolutional feature extraction with the modern vision architecture principles. These models allow making a comparison with the baseline CNN and assessing the usefulness of the pretrained CNN representations for galaxy morphology classification [17].

Figure 7 demonstrates the transfer-learning CNN backbone family used in this study. The representative ResNet, EfficientNet, and ConvNeXt/ConvNeXtV2 branches are followed by the global average pooling and the task-specific classification head.

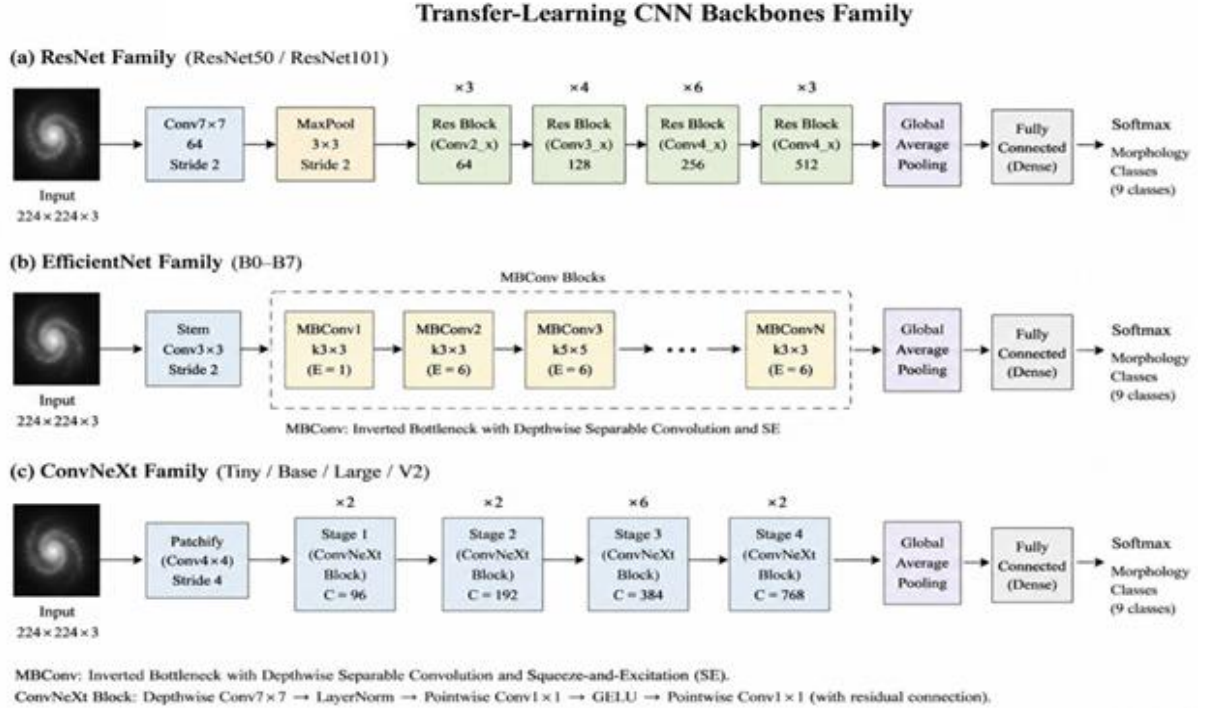


Figure 7. Transfer-learning CNN backbone family adapted for galaxy morphology classification.

3.4.3. Transformer and Hybrid Attention Family

The transformer and hybrid attention family was included in order to assess if the attention-based models can improve the representation of global and long-range spatial relationships in galaxy images. Since the CNN models allow capturing the local structures effectively, there is a need for broader contextual analysis of some galaxy morphology patterns beyond local neighborhoods. Transformer models and hybrid CNN-attention models are prominent in the field of galaxy morphology classification due to the ability of modeling dependencies beyond local neighborhoods [12].

This family includes the representative architectures such as Swin Transformer, MaxViT, FocalNet, and EVA-02. These models were chosen in order to analyze their different mechanisms for the integration of the local feature extraction with broader contextual reasoning [16].

The core architectural principle is the self-attention which calculates the relationships among the query, key, and value representations derived from image regions or tokens. The scaled dot-product attention is defined as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}}) V \quad (11)$$

where Q, K, and V represent the query, key, and value matrices, and d_k is the key dimension of the key.

In the Swin Transformer branch, the input image is divided into patches and converted into the token embeddings. The attention is applied locally to the patches using shifted-window attention in order to allow the exchange of information between neighboring regions, which leads to the hierarchical architecture.

In MaxViT, the local relationships within the regions are captured using the convolutional MBConv blocks, while the attention mechanisms are used to capture the broader relationships across the entire image.

FocalNet and EVA-02 are included as the stronger attention-based branches. The FocalNet uses the focal modulation in order to facilitate the feature interaction, while EVA-02 is a high-capacity transformer-style feature extractor. These models allow examining the improvement of the complex galaxy structure modeling using the broader attention-based representation.

Figure 8 demonstrates the transformer and hybrid attention family used in this study. The representative Swin Transformer, MaxViT, and FocalNet/EVA-02 branches are followed by the feature aggregation and the task-specific classification head.

Transformer and Hybrid Attention Family

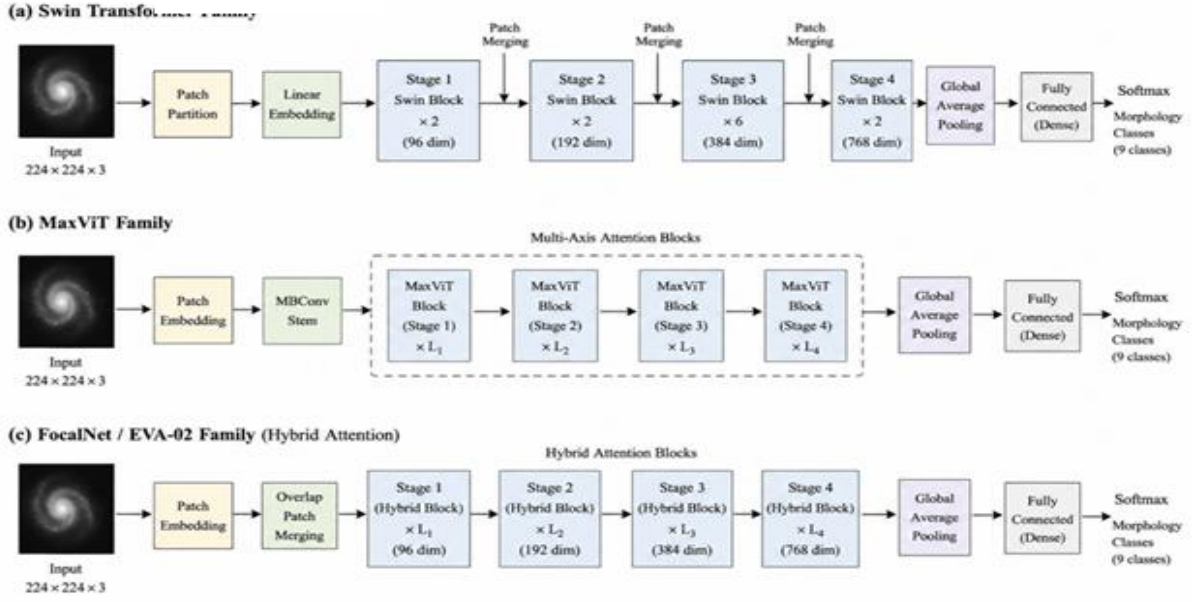


Figure 8. Transformers and hybrid attention architecture family used for broader spatial morphology reasoning.

3.4.4. YOLO Classification Family

The YOLO classification family was included in order to evaluate the efficient image-level classification branch within the proposed framework. Though YOLO architectures are known mainly as the object detection systems, their classification variants can be used for image-level predictions if bounding-box localization is not necessary. In this study, the YOLO classification branch is used as the fast and high-capacity alternative to conventional classification backbones, especially for the high-confidence Smooth versus Featured/Disk stage.

This family was chosen because YOLO classification models provide the practical image-level prediction pipeline. Architecturally, the YOLO classification branch follows the compact image classification pipeline: initial convolutional stem followed by the sequence of backbone feature blocks (e.g., CSP/C2f-style stages). The extracted feature maps are aggregated by pooling and fed into the linear classification layer followed by softmax layer.

The feature extraction stage of the YOLO classification branch is defined as:

$$h = f_{YOLO}(x) \quad (12)$$

where x is the input galaxy image, f_{YOLO} is the YOLO classification backbone, and h is the extracted feature representation.

The class probability for class c is calculated using the softmax function:

$$\hat{y}_c = \frac{\exp(z_c)}{\sum_{j=1}^C \exp(z_j)} \quad (13)$$

where z_c is the logit for class c , C is the number of morphology classes, and $\hat{y}_c$ is the predicted probability for class c .

Figure 9 illustrates the YOLO classification family used in this study. The image-level YOLO classification pipeline consists of the stem layer, backbone blocks, feature aggregation, global average pooling, fully connected classification layer, and the softmax morphology output.

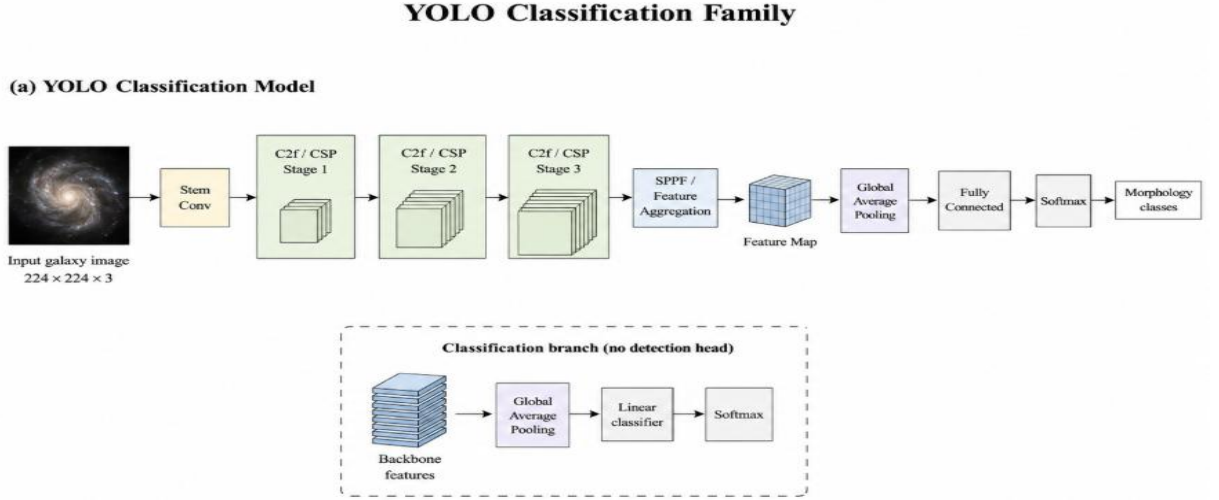


Figure 9. YOLO classification architecture family used as an efficient image-level morphology classifier

3.4.5. Ensemble and Hierarchical Classification Strategy

An ensemble and hierarchical classification strategy was implemented in order to integrate the outputs of different model families and expand the proposed framework from single-label classification towards the confidence-aware morphology analysis. Within the galaxy morphology research, a single final class may not be sufficient for the full description of galaxy image due to the visual ambiguity. Galaxy Zoo-style vote fractions are able to provide richer representation reflecting the agreement levels among classifiers and better describing uncertain or transitional morphologies than the hard label [1], [6].

The advantages of ensemble prediction are in the fact that different model families capture different information. The CNN models are able to capture the local morphology patterns, transfer-learning backbones provide the strong pretrained representation, transformers and hybrid attention models capture the broader spatial dependencies, and YOLO classification provides the efficient image-level prediction. Integration of these outputs supports a more stable prediction strategy than the single-model-family approach.

At the output level, the ensemble prediction is defined as follows:

$$\hat{y}_j = \sum_{m=1}^M w_m \hat{y}_{j,m} \quad (14)$$

where $\hat{y}_j$ is the final predicted output for morphology component j , $\hat{y}_{j,m}$ is the prediction of the model m for the same output, w_m is the corresponding model weight, and M is the number of models or model branches included in the ensemble.

The weights of the ensemble satisfy the following condition:

$$\sum_{m=1}^M w_m = 1 \quad (15)$$

Thus, the controlled integration of the complementary prediction sources is possible. For the hard-label classification branch, the fused prediction can be used for the class-level morphology decision-making. For the Galaxy Zoo branch, the same concept can be applied to the 37-output vote-fraction representation.

The final Galaxy Zoo vote-fraction output vector is:

$$\hat{\mathbf{y}} = [\hat{y}_1, \hat{y}_2, \dots, \hat{y}_{37}]$$

where the 37 outputs correspond to the Galaxy Zoo morphology hierarchy. This representation is more informative than a single label as it reflects the uncertainty and mixed morphological characteristics via vote-fraction values.

The hierarchical component of the strategy allows linking the broad high-confidence classification with the more detailed morphology prediction. For example, the high-confidence Smooth vs Featured/Disk decision can be considered as the initial morphology branch, while the full Galaxy Zoo output vector provides a more detailed hierarchical representation when vote-fractions are available [6].

Figure 10 illustrates the ensemble and hierarchical prediction strategy for this study. Predictions from diverse model families are integrated by the weighted fusion and connected to the broader 37-output Galaxy Zoo morphology prediction branch.

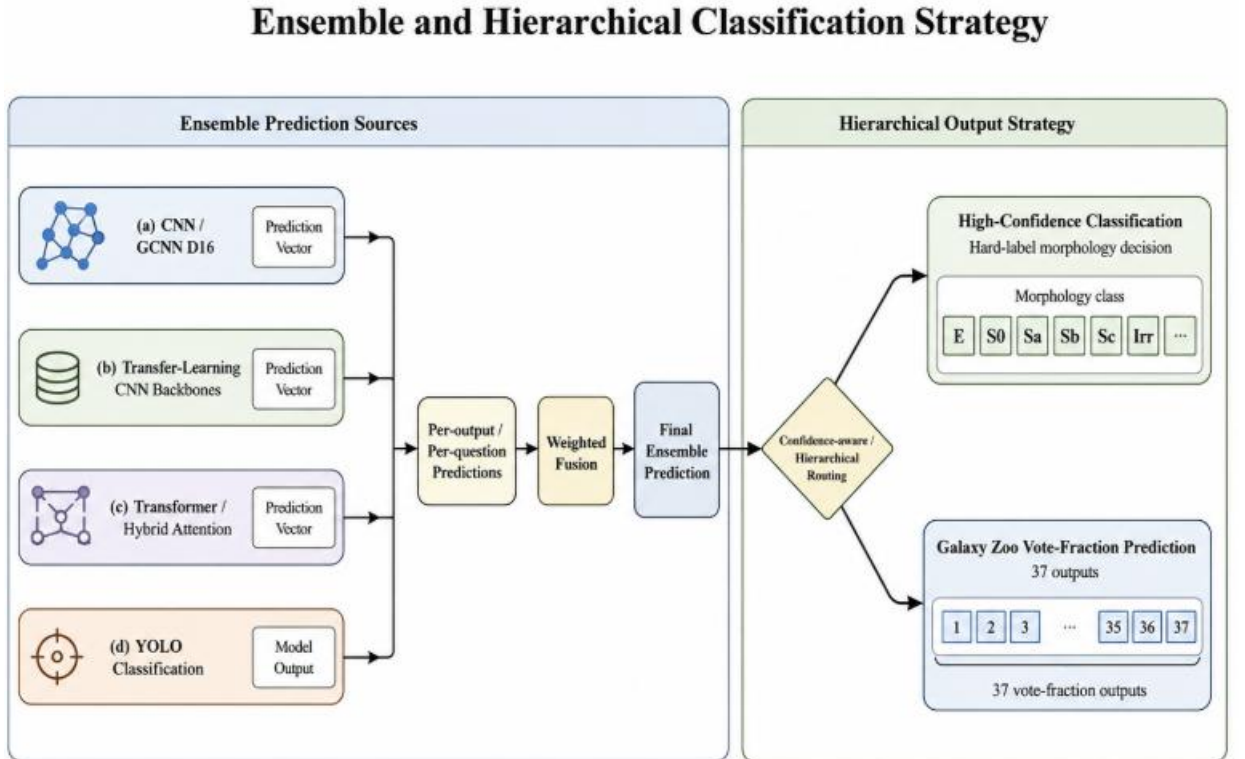


Figure 10. Ensemble and hierarchical prediction strategy connecting model families with 37-output Galaxy Zoo morphology prediction.

All together, these model families represent the classification architecture layer of the proposed framework. The CNN and GCNN branches provide the local and symmetry-aware feature extraction, the

transfer-learning CNN backbones provide the pretrained representation, transformers and hybrid attention models capture the broader spatial relationships, and the YOLO classification branch allows efficient high-confidence image-level prediction. The ensemble and hierarchical strategy integrate these outputs into the broader Galaxy Zoo vote-fraction representation, allowing to move from controlled hard-label galaxy morphology classification to confidence-aware hierarchical morphology analysis.

3.5. Training and Evaluation

The training and evaluation processes are designed in accordance with the role of each prediction branch within the framework. In the case of hard-label classification, models are trained using the Galaxy10 DECaLS training subset and validated using the validation subset, while the held-out test subset is reserved for the final evaluation. For the Galaxy Zoo branch, the labeled data are split into training, validation, and internal test subsets, while the official unlabeled Kaggle test images are reserved for the final prediction submission [10].

Training is performed using supervised learning in accordance with the target structure of each task. The Galaxy10 DECaLS branch solves the multi-class morphology classification task, while the high-confidence Smooth vs Featured branch is considered as the binary classification task. The Galaxy Zoo branch is considered as the continuous multi-output prediction task since each galaxy image has 37 vote-fraction targets rather than a single class label.

For hard-label classification, models are trained to predict the single morphology class per galaxy image using the cross-entropy loss:

$$L_{CE} = -1 \times \sum_{c=1}^C y_c \log(\hat{y}_c) \quad (16)$$

where C is the number of morphology classes, y_c is the ground-truth label, and $\hat{y}_c$ is the predicted probability for class c.

For the Galaxy Zoo vote-fraction branch, the task is considered as the continuous multi-output prediction task because of 37 vote-fraction targets. Therefore, the evaluation includes regression-oriented metrics such as RMSE and MAE, as well as per-output analysis of the predicted vote-fractions.

The performance of classification is evaluated using accuracy, precision, recall, F1-score, and confusion matrix. In the case of the 37-output Galaxy Zoo branch, RMSE and MAE are used to measure vote-fraction prediction quality.

In addition to the standard analysis for assessing the performance of the model, reliability analysis was used. The robustness of the algorithm was assessed through prediction behavior on image manipulation, uncertainty was considered to provide confidence-based interpretation for ambiguous images of galaxies, and Grad-CAM was used to highlight the image areas that have the most influence on model predictions. This assessment of reliability is essential for the main goal of the research, which is not only accurate classification of the galaxy morphology but also reliable predictions [3].

4. Experimental Results and Discussion

This section is devoted to experimental results and discussion concerning the proposed galaxy morphology framework. Subsections below will cover the outputs from the preprocessing stage, the hard-label classification evaluation, the Galaxy Zoo vote-fraction prediction evaluation, and the reliability-oriented evaluation.

The results in this section are organized according to the main components of the proposed framework. First, the preprocessing outputs are presented to verify that the data used in the experiments are consistent with the methodology described in Section 3. Second, the hard-label Galaxy10 DECaLS classification results are reported across the evaluated model families, followed by a detailed analysis of the selected ConvNeXtV2 transfer-learning CNN model. Third, the high-confidence Smooth vs Featured classification stage is evaluated as a binary morphology-routing task. Fourth, the Galaxy Zoo 37-output vote-fraction prediction branch is assessed using RMSE, MAE, per-output analysis, first-gate analysis, and representative lowest-error and highest-error examples. Finally, reliability-oriented evidence is presented through robustness, confidence, and visual explanation analyses.

The results are reported using numerical metrics, confusion matrices, and corresponding visual outputs, while the discussion focuses on model performance, class-level errors, robustness, uncertainty-related behavior, and explainability evidence.

4.1. Preprocessing Results

Preprocessing will be described first to prove that the experiments comply with the methodology. The hard-label classification branch uses the Galaxy10 DECaLS class labels, while the vote-fraction prediction branch preserves continuous vote fractions from Galaxy Zoo. The Smooth vs Featured classification stage can be regarded as a high-confidence binary decision based on the first question from Galaxy Zoo.

Table 4. Preprocessing outputs used in the framework.

Data component	Label type	Preprocessing result
Galaxy10 DECaLS	Hard labels	Official ten-class classification and reduced diagnostic analysis
Galaxy Zoo Challenge	37 vote fractions	Continuous hierarchical prediction and first-gate analysis
Smooth vs Featured subset	High-confidence binary labels	7905 training, 1694 validation, 1694 held-out test samples

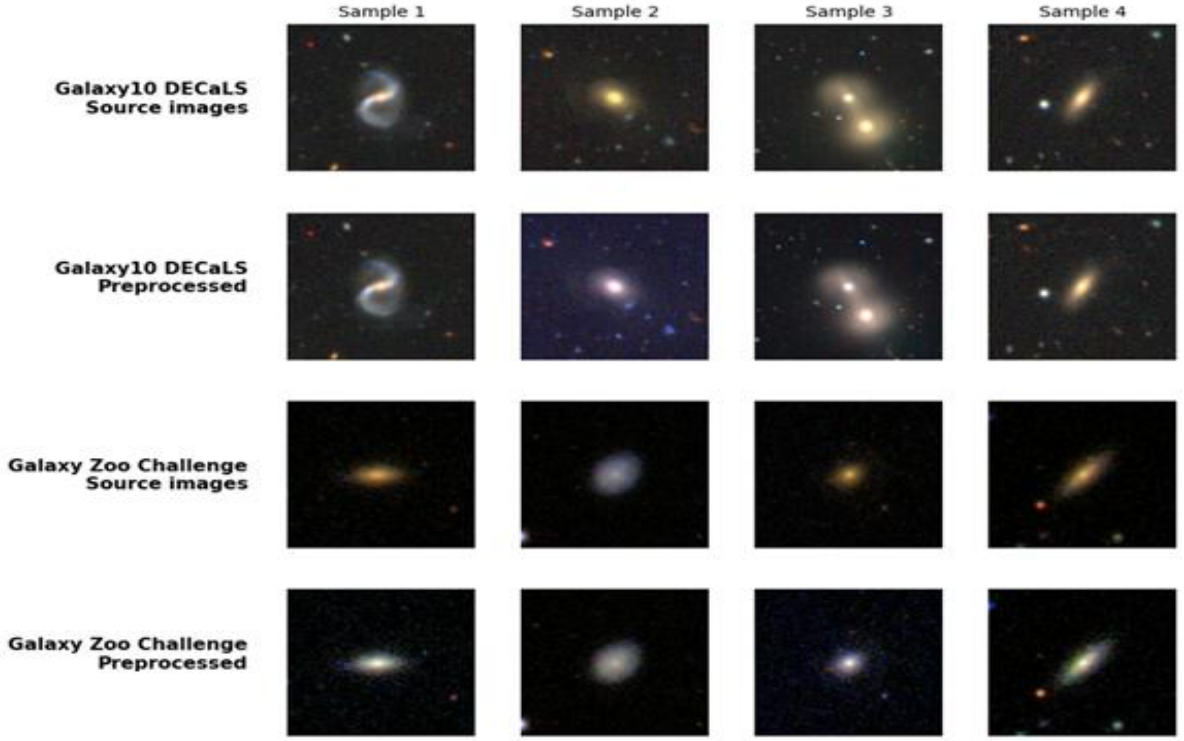


Figure 11. Representative source and preprocessed galaxy images from the Galaxy10 DECaLS and Galaxy Zoo datasets

4.2. Hard-Label Classification Results Across Model Families

Hard-label classification results of the examined families of models are provided in this section. It should be noted that the results are presented not on an individual experiment basis but rather for each family of models. This is done because section 3 specifies what is the methodological role of family models, whereas section 4 highlights best-performing examples of each family and provides detailed analysis of the final relevant model.

Table 5. Hard-label classification performance across the evaluated model families.

Model family	Representative experiment	Task setting	Accuracy / Recall / F1-score	Evaluation status	Interpretation
CNN / GCNN family	Baseline CNN and GCNN D16	Galaxy10 full 10-class	Validation accuracy = 0.886466; Macro Recall = 0.876281; Macro F1 = 0.878177	Validation evidence	Provides the baseline and symmetry-aware reference for hard-label galaxy morphology classification.
Transfer-learning CNN family	ConvNeXtV2 transfer-learning CNN model	Galaxy10 full 10-class	Test accuracy = 0.915821; Macro Recall = 0.908692; Macro F1 = 0.907277	Held-out test evidence	Provided the highest full ten-class test performance in this comparison and was used for the detailed per-class and confusion-matrix analysis.
Transformer / hybrid	Swin Large 384 attention model	Galaxy10 full 10-class	Validation accuracy =	Validation evidence	Shows the contribution of attention-based

attention family			0.884962; Macro Recall = 0.861231; Macro F1 = 0.862763		representations in the full ten-class hard-label classification setting.
YOLO11 classification family	YOLO11x-cls and YOLO11 ensemble classifiers	Galaxy10 full 10-class	Validation Top-1 = 0.893233; weak-class recall mean = 0.753126; Macro F1 = 0.878678	Validation evidence	Confirms that YOLO11 was evaluated as an independent hard-label classification family on the full ten-class Galaxy10 task.
Ensemble and fusion strategy	YOLO11 + FocalNet reduced-class fusion	Reduced 7-class validation-clean diagnostic	Accuracy = 0.980296; Macro Recall = 0.971904; Macro F1 = 0.970908	Validation-only diagnostic evidence	Shows that model-family diversity improved the reduced-class diagnostic result, while this evidence remains separate from the original full ten-class task.

Table 5 demonstrates that the transfer-learning CNN family provided the best overall performance on the full ten-class Galaxy10 test. At the same time, the YOLO11 family also produced good hard-label results, especially in the reduced seven-class diagnostic test. The transformer/hybrid and fusion families helped to ensure the benefits of diversity of approaches and class-level ambiguity analysis. Based on the results obtained on the full ten-class test, the ConvNeXtV2 transfer-learning CNN model was selected to carry out the detailed per-class and confusion-matrix analysis in the following subsection.

4.2.1. Detailed Analysis of the ConvNeXtV2 Transfer-Learning CNN Model

Base evaluation. Having examined the hard-label classification families, the ConvNeXtV2 transfer-learning CNN model was analyzed in detail due to its superior performance on the full ten-class Galaxy10 DECaLS test. Base evaluation, confusion-matrix analysis, examples of correct/incorrect classification, interpretability evidence, robustness test, and uncertainty estimation are included in the analysis.

Table 6. Per-class classification report for the ConvNeXtV2 transfer-learning CNN model on the Galaxy10 DECaLS ten-class test set.

<i>Class</i>	<i>Precision</i>	<i>Recall</i>	<i>F1-score</i>	<i>Support</i>
Class 0	0.759	0.778	0.768	162
Class 1	0.967	0.942	0.954	278
Class 2	0.989	0.945	0.966	397
Class 3	0.986	0.961	0.973	304
Class 4	0.830	0.880	0.854	50
Class 5	0.976	0.915	0.945	307
Class 6	0.808	0.887	0.846	275
Class 7	0.811	0.860	0.835	394
Class 8	0.971	0.958	0.965	213

Class 9	0.971	0.961	0.966	281
Accuracy	—	—	0.916	2661
Macro Avg	0.907	0.909	0.907	2661
Weighted Avg	0.920	0.916	0.917	2661

Confusion matrix analysis. Confusion matrix in Figure 12 demonstrates a strong diagonal pattern implying that most of the test samples were classified correctly. However, the remaining errors are not evenly distributed across all the classes; instead, they cluster around visually ambiguous classes, namely Class 7 with Class 6, Class 0 with Class 7, and Class 6 with Class 7. Thus, further diagnostic analysis could be conducted to explore weak classes.

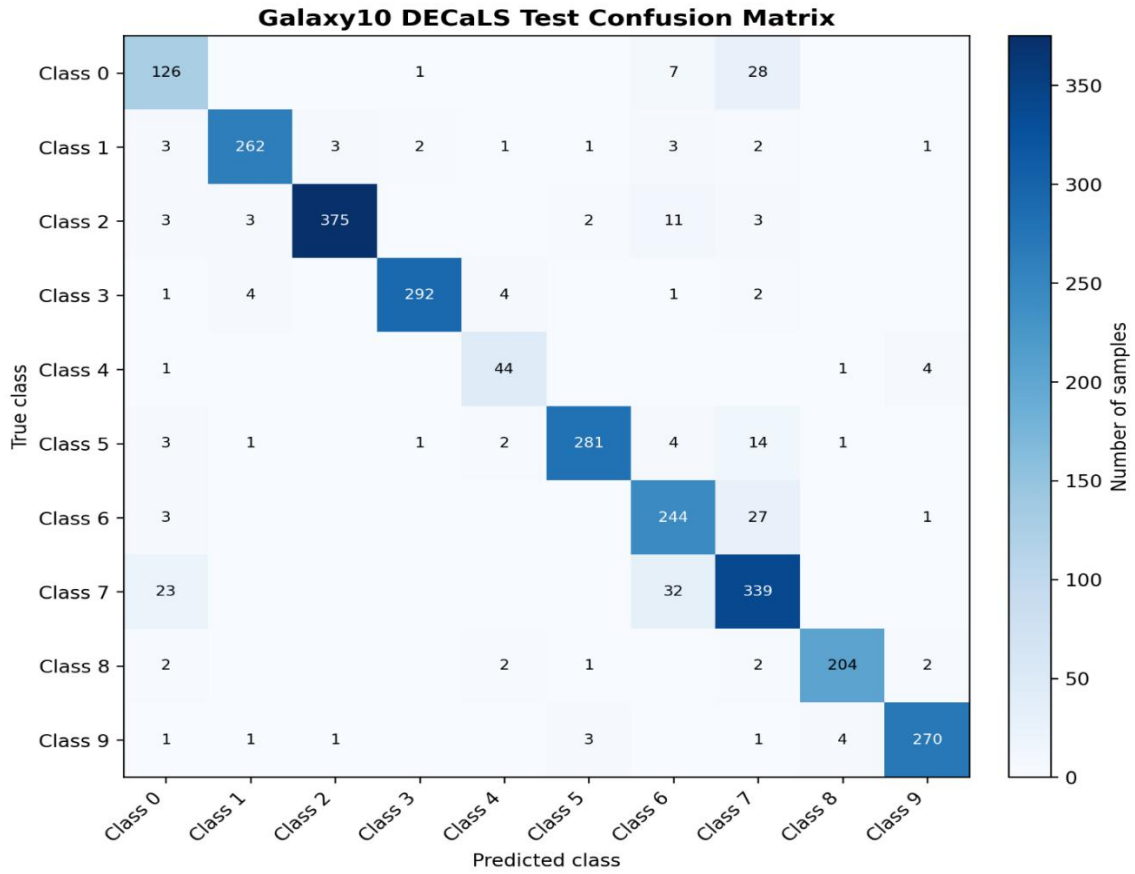


Figure 12. Galaxy10 DECaLS ten-class confusion matrix on the held-out test set.

Qualitative correct/incorrect analysis. To complement the numeric evaluation, representative correct and incorrect predictions were analyzed from the held-out Galaxy10 DECaLS test set. Correct examples illustrate the cases when the model made predictions about the expected morphology class with high confidence, while incorrect examples illustrate visually ambiguous cases when the predicted class differs from the true class label. Figure 13 provides representative examples of correct and incorrect prediction.

As can be seen from the examples below, the confusion-matrix pattern is confirmed as the most errors occur in classes having visually similar structure.

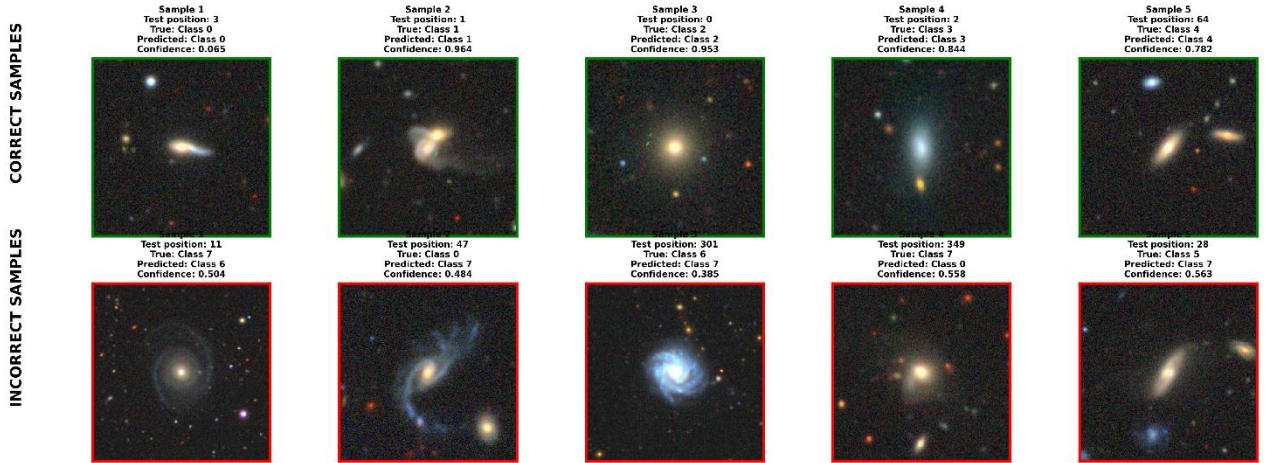


Figure 13. Representative correct and incorrect Galaxy10 DECaLS predictions from the ConvNeXtV2 transfer-learning CNN model.

Interpretability. Interpretability analysis was conducted to check whether the model attended to important galaxy regions when making predictions. Visual explanations based on Grad-CAM revealed the regions which contributed the most to the classification decision. This step is important since high classification performance does not mean that the model takes into account scientifically meaningful morphological features. Figure 14 illustrates representative Grad-CAM visualizations for the selected Galaxy10 DECaLS samples. The heatmap overlay represents the parts of the image affecting model predictions the most (e.g., central brightening, disk-like structure, or elongated shape).

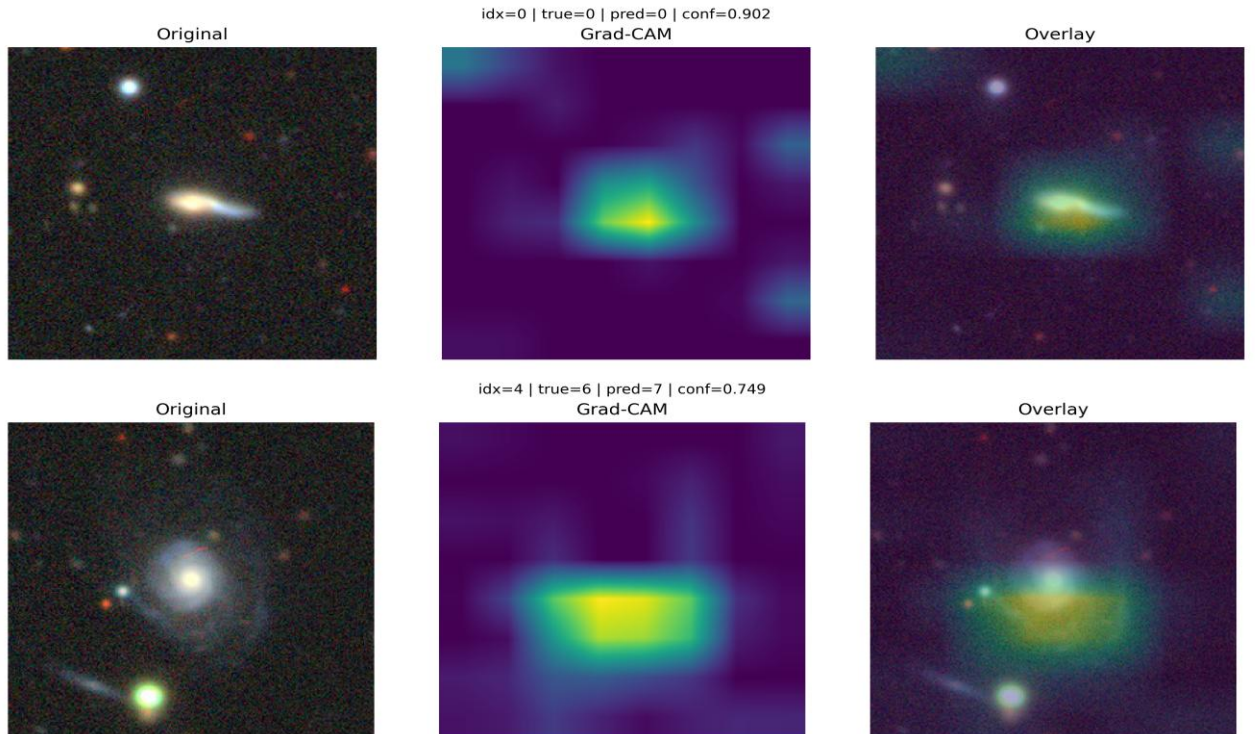


Figure 14. Grad-CAM visual explanations for selected Galaxy10 DECaLS predictions from the ConvNeXtV2 transfer-learning CNN model.

Robustness. Robustness analysis was carried out to see how the model behaved when there were controlled image perturbations, namely noise, changes of brightness, and blur. This step revealed whether prediction performance remained steady when input images were changed in certain ways.

Figure 15 summarizes the results of the robustness evaluation under image perturbations. The most significant degradation occurred with noise and blur, thus, implying the importance of good image quality for hard-label galaxy morphology classification.

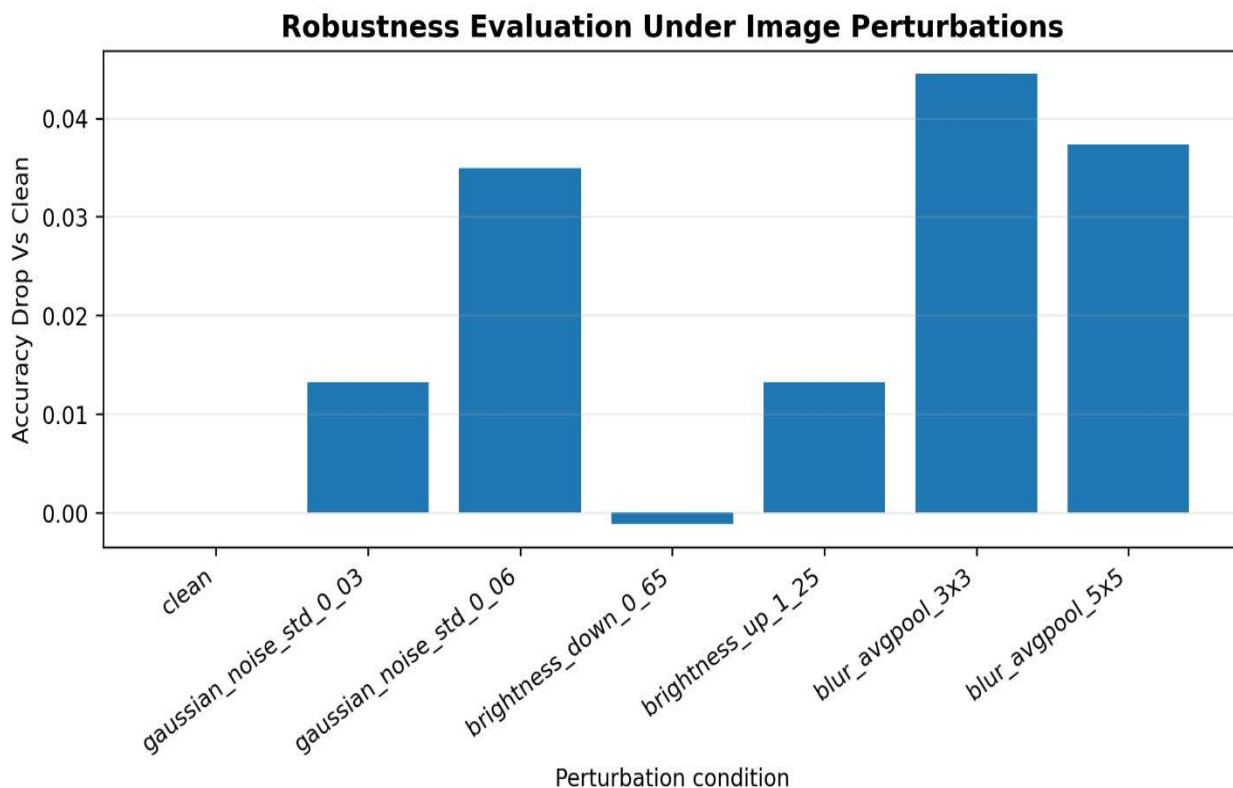


Figure 15. Robustness evaluation under controlled image perturbations.

Uncertainty estimation. Uncertainty estimation was performed to compare the confidence profiles of correct and incorrect predictions and see if incorrect predictions had lower confidence than correct predictions.

Figure 16 provides the uncertainty distributions for the correct and incorrect predictions. The distribution supported the reliability-oriented component of the framework by showing that the incorrect predictions tended to be less confident than correct ones.

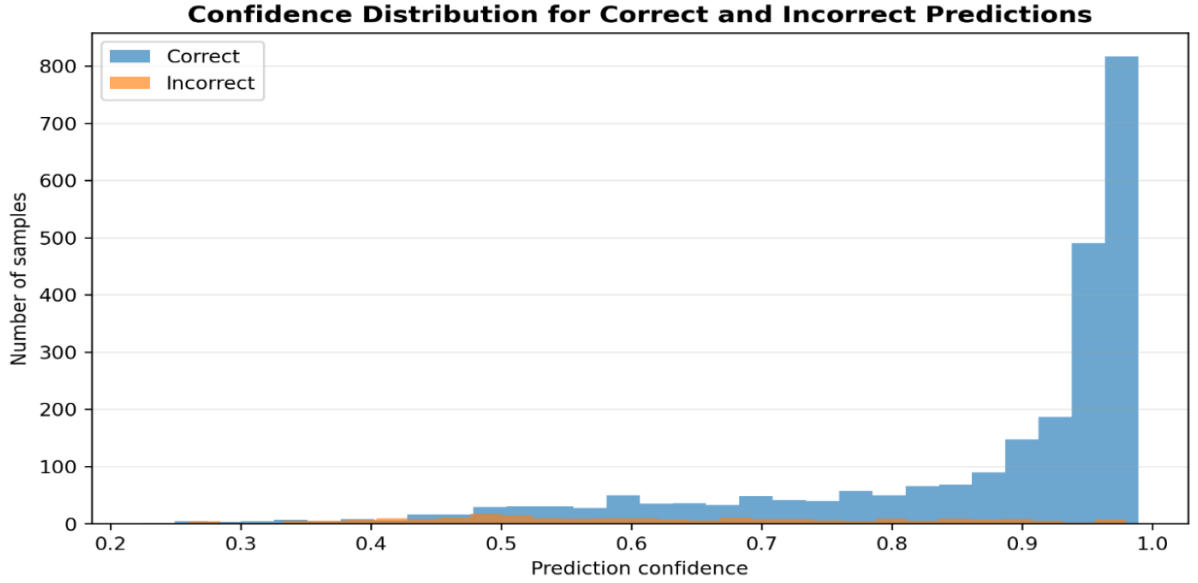


Figure 16. Uncertainty distribution for correct and incorrect Galaxy10 DECaLS predictions.

Thus, the full ten-class hard-label classification on Galaxy10 DECaLS dataset represented the most difficult full-class classification setting in this section. The official evaluation on the held-out test achieved 91.58% accuracy and macro F1-score of 0.9073, meaning good performance on the full ten-class morphology task. The per-class results, the confusion matrix, examples of correct and incorrect classification, robustness analysis, and uncertainty-based evaluation proved reliable performance for most classes, while the rest of the errors occurred in visually similar or weakly separated morphology classes. This made the full ten-class Galaxy10 DECaLS experiment the main baseline for evaluating the proposed framework in realistic multi-class morphology classification setting.

4.3. High-Confidence Smooth vs Featured Classification

This subsection provides the detailed analysis of the high-confidence Smooth vs Featured classification stage, which evaluates the first-level morphology routing component of the proposed framework. Within this stage, the YOLO classification family described in Section 3 was used, specifically YOLO11x, as the image-level classifier for the binary Smooth vs Featured task. Galaxies with high Galaxy Zoo first-gate evidence were assigned to either the Smooth or Featured class.

The binary subset was created based on a confidence threshold of 0.90. Galaxies with a Smooth vote fraction of 0.90 or higher were assigned to the Smooth class, while galaxies with a Featured/Disk vote fraction of 0.90 or higher were assigned to the Featured class. The final evaluation was conducted on the fixed held-out test set after selecting the model protocol on the validation set only.

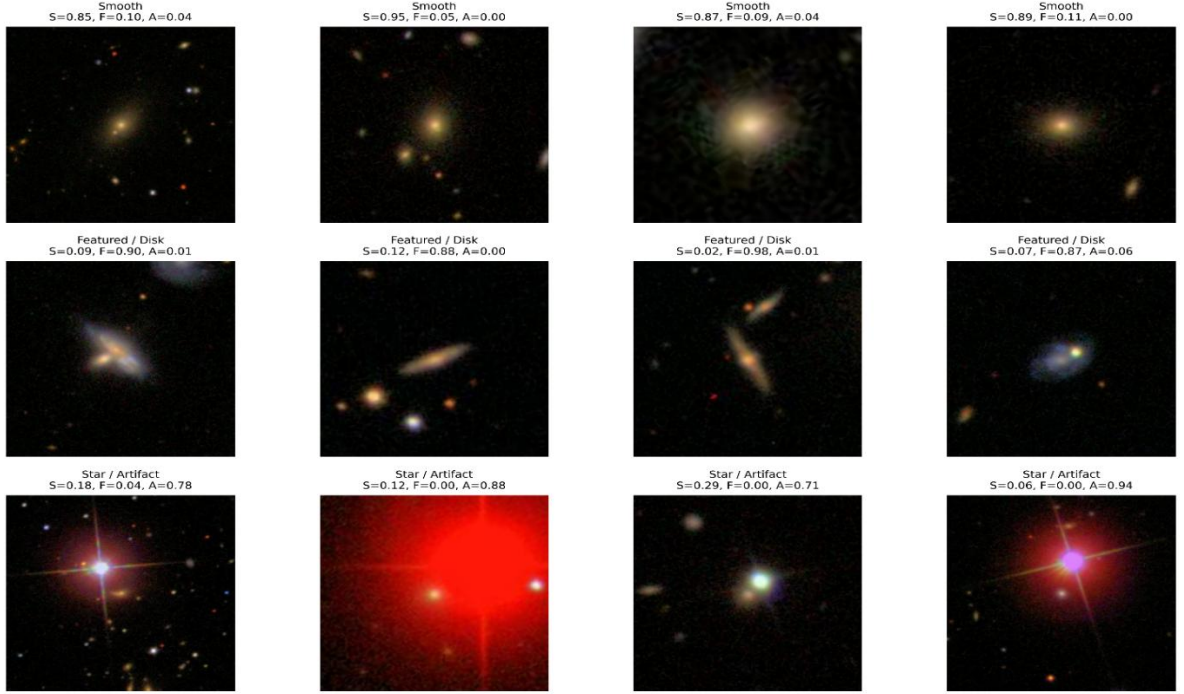


Figure 17. Galaxy Zoo first-gate vote-fraction examples used to construct the high-confidence Smooth vs Featured binary classification task.

Base Evaluation. Table 7 presents the official evaluation results on the held-out test for the high-confidence Smooth vs Featured classifier. The model achieved an overall accuracy of 99.1736% and the macro F1-score of 0.988126 on 1,694 test galaxies.

Table 7. Per-class classification report for the YOLO11x high-confidence Smooth vs Featured classifier on the held-out test set.

Class	Precision	Recall	F1-score	Support
Featured	0.9962	0.9932	0.9947	1316
Smooth	0.9764	0.9868	0.9816	378
Accuracy	—	—	0.9917	1694
Macro avg	0.9863	0.9900	0.9881	1694
Weighted avg	0.9918	0.9917	0.9918	1694

The model achieved the accuracy of 99.1736% with a macro F1-score of 0.9881 on the held-out test. The Featured class got the F1-score of 0.9947, while the Smooth class achieved the F1-score of 0.9816, thus providing good performance for both binary morphology classes.

Confusion Matrix Analysis. Confusion matrix was used to explore the distribution of classification errors between the Smooth and Featured classes, and thus prove whether errors occurred in one of the classes only or were evenly distributed among both binary classes.

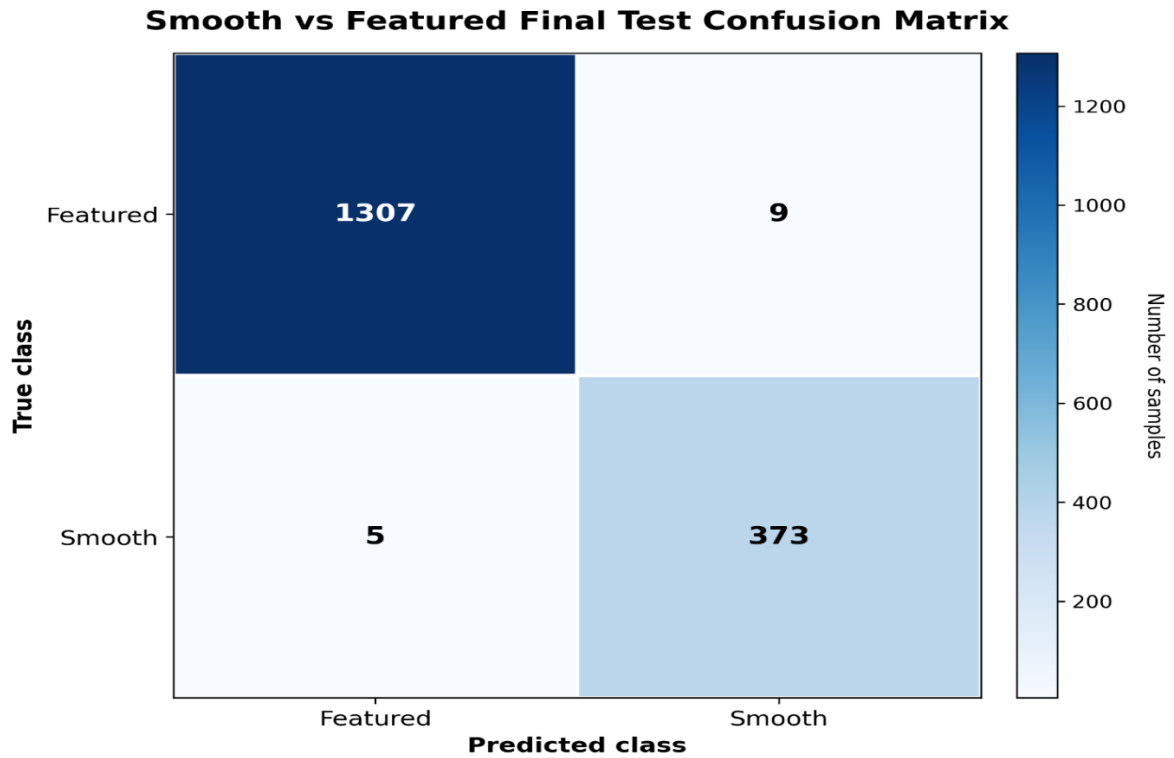


Figure 18. Official held-out test confusion matrix for the high-confidence Smooth vs Featured binary classification stage.

As can be seen in Figure 18, the model classified 1,307 Featured galaxies and 373 Smooth galaxies correctly. The number of errors was rather limited as only 9 Featured galaxies were predicted as Smooth and 5 Smooth galaxies were predicted as Featured. Therefore, the results confirmed the overall test performance provided in Table 8 and proved that the binary classifier maintained an even error distribution between the two classes.

Qualitative Analysis. Qualitative examples were used to provide representative correct and incorrect predictions, and thus find out whether the model decisions corresponded to visible galaxy morphology and whether the remaining errors corresponded to the visually challenging cases.

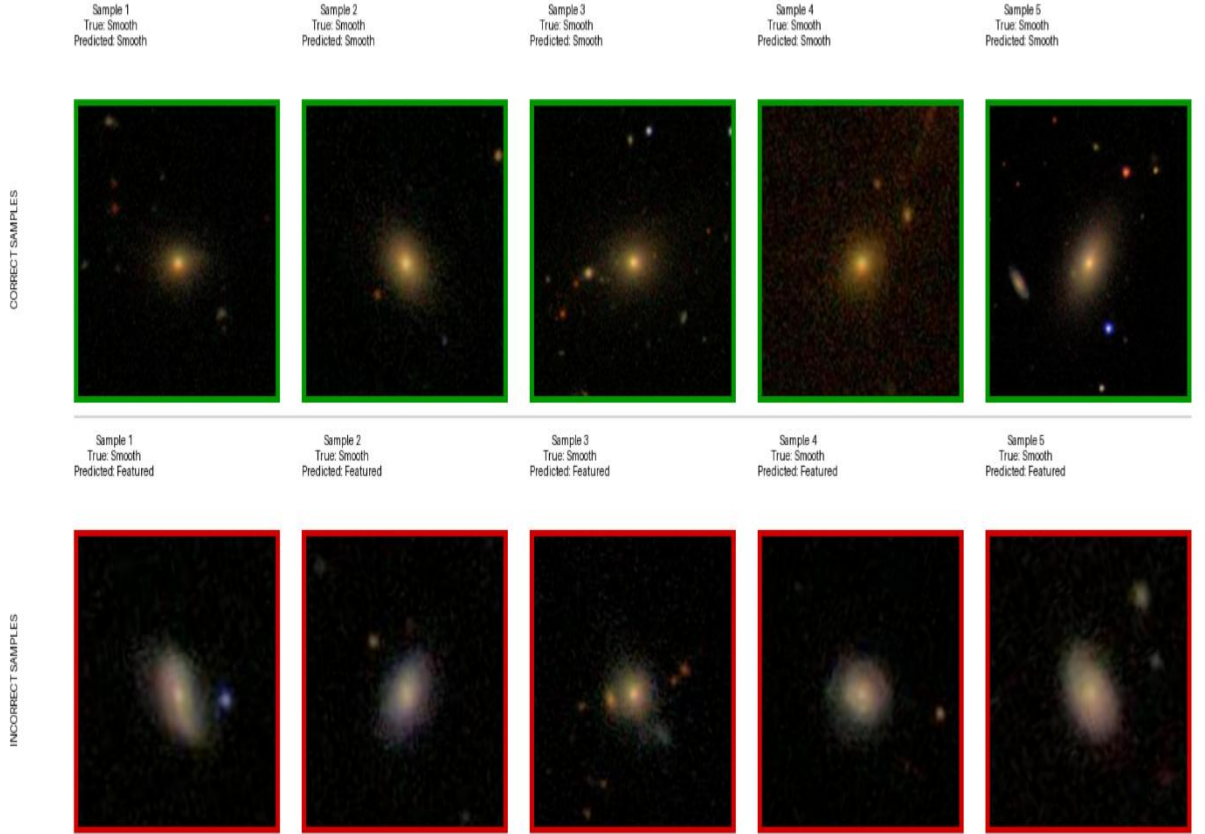


Figure 19. Representative correct and incorrect predictions from the high-confidence Smooth vs Featured classification stage.

Figure 19 showed that correct examples had rather clear morphology patterns corresponding to their class, while incorrect examples looked more ambiguous in terms of galaxy morphology, which had weak structure, low contrast, or smooth light distribution and made the Smooth vs Featured differentiation hard. Thus, the qualitative examples proved the quantitative results, showing that the remaining errors were rather limited and happened in challenging visual cases.

Selective Confidence Analysis. Figure 18 explored the classifier behavior under the confidence threshold, highlighting the ability of the framework to select those predictions which could be accepted with higher reliability.

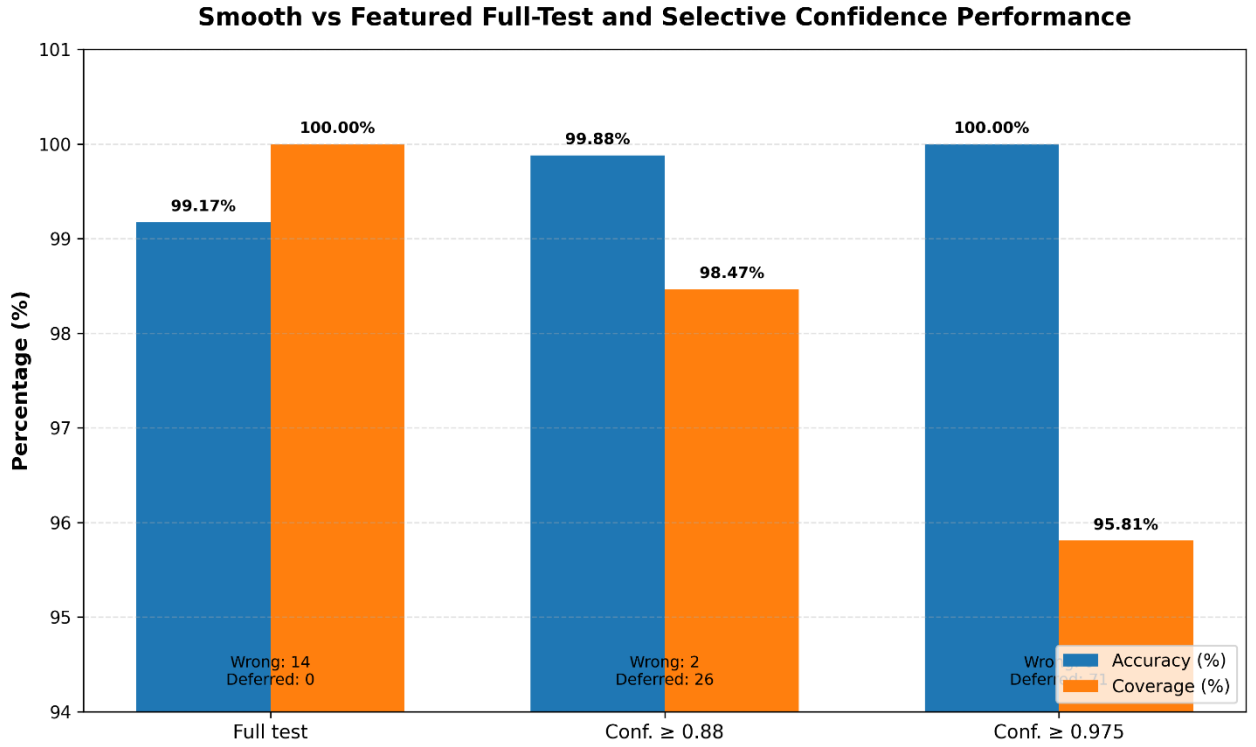


Figure 20. Full-test and selective-confidence performance of the Smooth vs Featured classification stage.

When there was no selective filtering, the model achieved 99.1736% accuracy at full test coverage. If the confidence threshold was increased to 0.88, the model accepted predictions with 99.8801% accuracy and 98.4652% coverage. When the threshold became even stricter at 0.975, the model accepted predictions with 100.0000% accuracy and 95.8087% coverage.

Thus, there was a trade-off between the coverage and reliability of accepted predictions since stricter confidence threshold led to lower number of accepted samples but higher prediction reliability.

Robustness Analysis. To check the stability of the model, the held-out test images were subjected to controlled image perturbations, including brightness change, contrast change, blur, and Gaussian noise. Figure 21 compared the accuracy of the model under these conditions and the accuracy in the clean image reference condition.

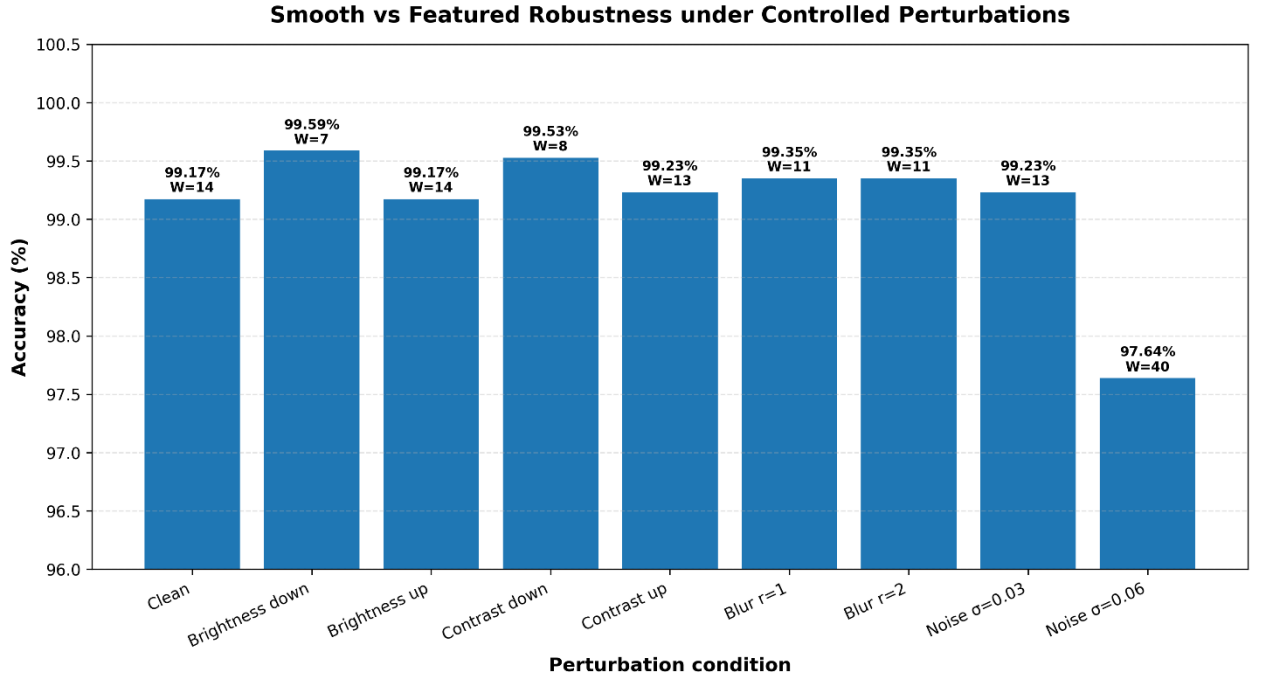


Figure 21. Robustness of the Smooth vs Featured classifier under controlled image perturbations.

The model maintained high accuracy under most of the perturbations, remaining at or above 99% for brightness, contrast, blur, and mild Gaussian noise. However, stronger Gaussian noise led to significant degradation and dropped the accuracy to 97.6387%, thus suggesting that the model was stable under most of the moderate image variations but sensitive to stronger noise that affected faint morphology cues.

Overall, the high-confidence Smooth vs Featured classification stage proved its strong held-out test performance, achieving 99.1736% accuracy and a macro F1-score of 0.9881. Confusion matrix and qualitative examples proved that errors were rather limited and happened in visually challenging cases. Selective confidence analysis showed that prediction reliability could be increased by accepting only higher-confidence cases, while robustness analysis showed stable performance under most of the moderate image perturbations. Thus, these results proved that the stage could be used as a reliable first-level morphology routing component of the proposed framework.

4.4. Galaxy Zoo Vote-Fraction Prediction Results

This subsection evaluates the Galaxy Zoo vote-fraction prediction branch of the proposed framework. Unlike the hard-label classification stage, this branch predicts 37 continuous vote-fraction outputs for each galaxy image, making the morphology prediction task more detailed and challenging. This branch belongs to the transfer-learning CNN backbone family described in Section 3, with ConvNeXtV2 Large used as the main held-out test architecture in the study evaluation. Post-deadline Kaggle scores are considered additional benchmarking evidence and are not treated as the primary evaluation protocol.

The purpose of this experiment is to find out if the framework can learn the distribution of human morphology votes beyond simple hard-label classification. This point is important since Galaxy Zoo annotations may represent some uncertainty or partial agreement which cannot be properly reflected in a single-class label.

Post-deadline Kaggle scoring evidence is reported later in Table 11 as an additional external benchmark for this vote-fraction branch, but it is not treated as the primary evaluation protocol of the study.

Base Evaluation. Table 8 provided the official results on the held-out test for the ConvNeXtV2 Large architecture on the Galaxy Zoo 37-output vote-fraction prediction task. RMSE and MAE measured prediction errors, and lower values meant better agreement with the target vote fractions.

Table 8. Official Galaxy Zoo 37-output vote-fraction prediction results on the held-out test set.

Architecture	Output type	Evaluation set	RMSE	MAE	First-gate accuracy
ConvNeXtV2 Large	37 continuous vote fractions	Held-out test	0.074912	0.042315	0.887626

The model achieved an RMSE of 0.074912 and an MAE of 0.042315. The first-gate accuracy of 0.887626 indicated that continuous predictions still had useful high-level morphology information.

Per-output Error Analysis. Figure 22 illustrated the RMSE values for the 37 Galaxy Zoo vote-fraction outputs and thus checked whether the overall RMSE was uniform across the hierarchy or affected by morphology questions.

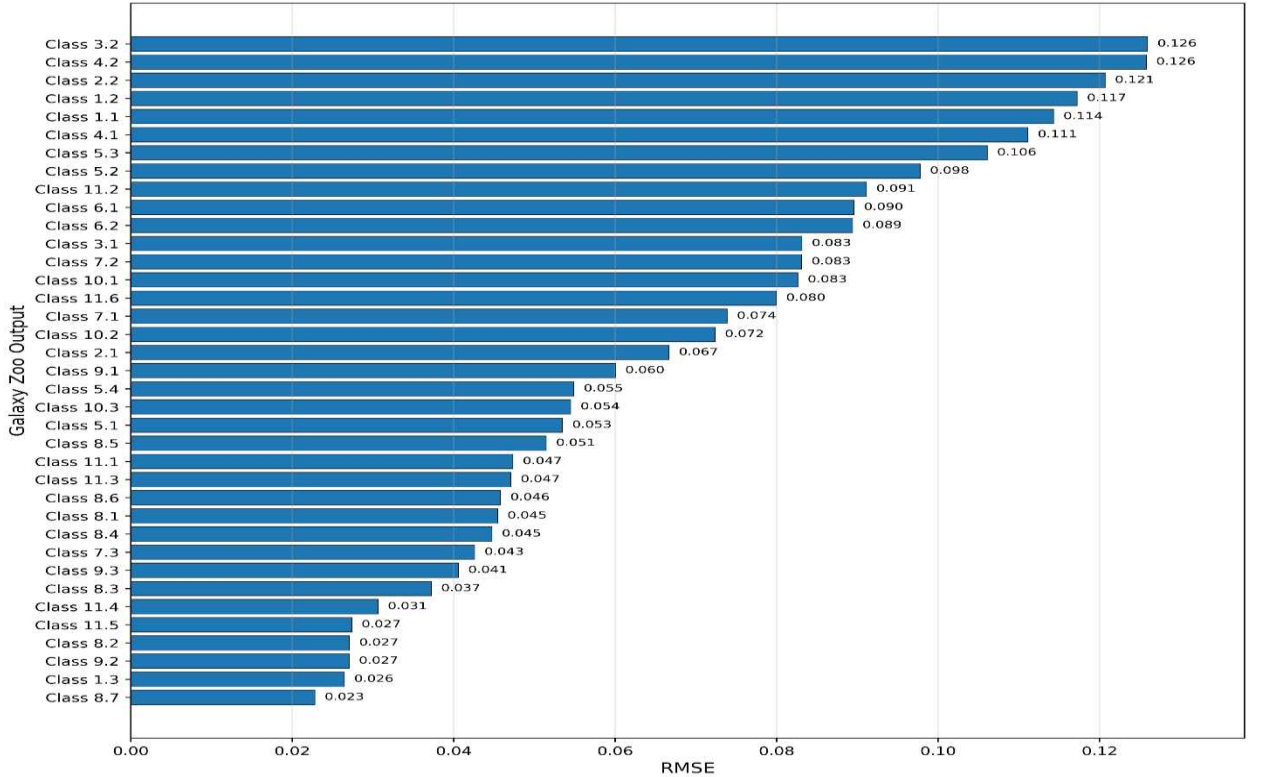


Figure 22. Per-output RMSE values for the Galaxy Zoo 37-output vote-fraction prediction task.

The results showed that prediction error was not equal across the 37 outputs. Outputs with clearer or more common morphology responses had lower error, while outputs with subtler or less frequent features had higher error. Therefore, it meant that the model achieved good overall performance with some detailed morphology attributes being harder to predict.

First-gate Analysis. First-gate analysis helped to check whether the 37-output vote-fraction model retained the main Galaxy Zoo morphology decision. The dominant first-gate category was determined from the predicted vote fractions and compared to the corresponding target category.

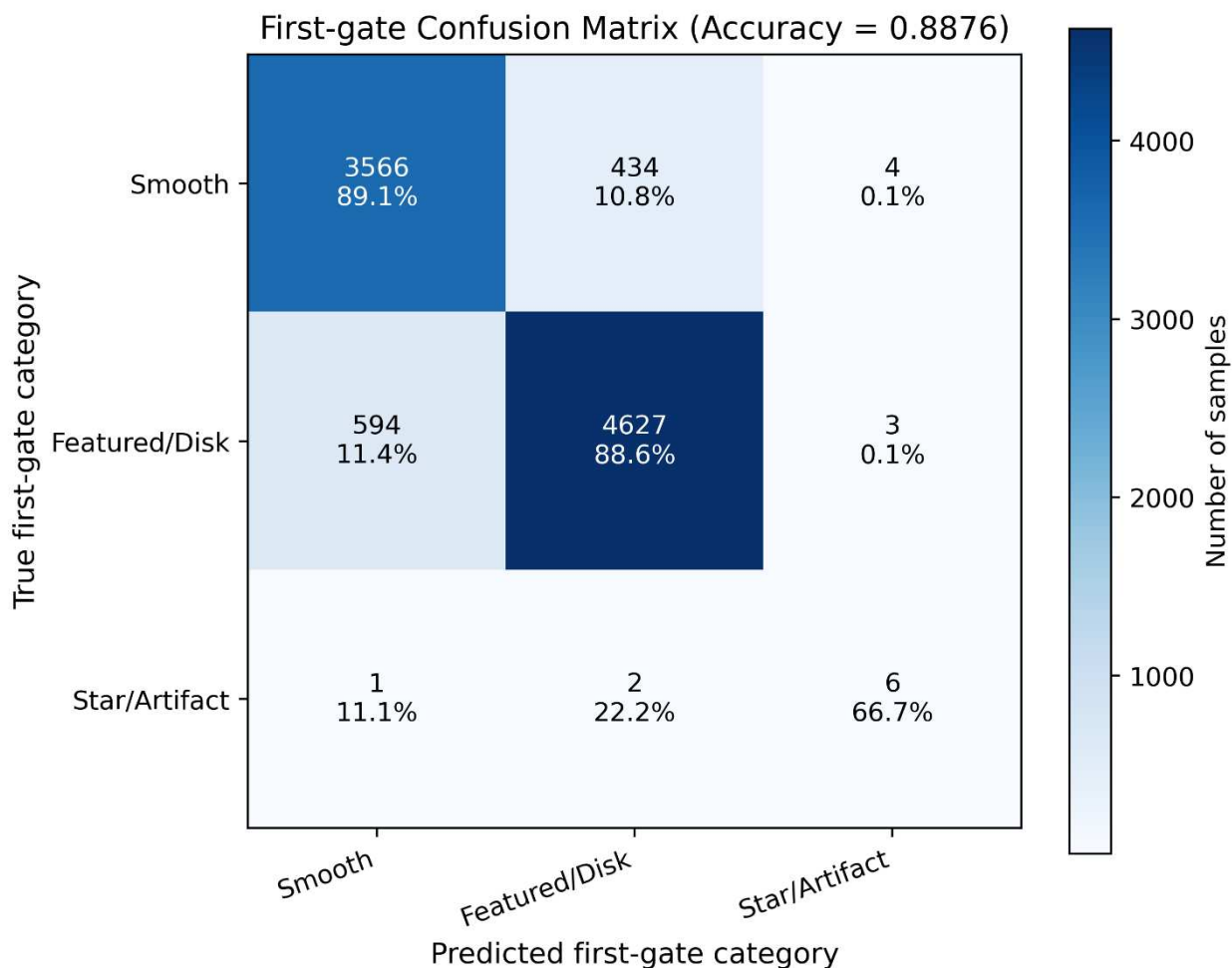


Figure 23. First-gate prediction analysis for the Galaxy Zoo 37-output vote-fraction model.

The model achieved a first-gate dominant-class agreement of 0.887626 on the held-out test set. This metric was used only as a secondary diagnostic to verify whether the continuous vote-fraction predictions preserved the main Galaxy Zoo morphology decision. Even though the model was trained to predict continuous vote fractions instead of single label, it managed to retain useful high-level morphology information. Most of the errors occurred when first-gate vote fractions were close or visually ambiguous.

Qualitative Best/Worst Prediction Examples. Representative best and worst prediction examples were selected based on sample-level prediction error to correlate RMSE results with visually galaxy cases and prove when the 37-output model agreed more or less with the target vote fractions.

Lowest-error examples



Highest-error examples



Figure 24. Representative lowest-error and highest-error Galaxy Zoo vote-fraction prediction examples.

Thus, the best examples showed vote-fraction patterns corresponding to the target morphology response, while the worst examples showed difficult cases with visually ambiguous or mixed morphology resulting in bigger differences between the predicted and target vote fractions. Therefore, the qualitative analysis confirmed the quantitative results, showing that higher errors occurred in difficult morphology cases rather than in randomly chosen cases.

Overall, the Galaxy Zoo 37-output vote-fraction experiment proved that the framework could predict continuous morphology responses with strong agreement to the target vote fractions. The ConvNeXtV2 Large model achieved the RMSE of 0.074912 and the MAE of 0.042315 on the held-out test set, while the first-gate accuracy of 0.887626 indicated retention of useful high-level morphology information. The per-output and qualitative analysis showed that higher errors were primarily associated with more detailed or visually ambiguous morphology responses. Therefore, these results proved the application of the vote-fraction prediction as the component of the framework for modeling morphology uncertainty beyond single-label classification.

4.5. Reliability-Oriented Evaluation

Table 9 presented the main reliability evidence obtained from robustness testing, selective-confidence analysis, confidence/error diagnostics, and saliency-based visual explanations.

Table 9. Quantitative reliability evidence supporting the proposed galaxy morphology framework.

<i>Reliability component</i>	<i>Experiment / evidence source</i>	<i>Quantitative evidence</i>	<i>Interpretation</i>
Robustness under perturbations	Galaxy Zoo 37-output vote-fraction prediction	Clean RMSE = 0.074912; Gaussian noise RMSE = 0.095387	Gaussian noise produced the clearest degradation, showing that strong image noise affects continuous vote-fraction prediction.
First-gate stability	Galaxy Zoo 37-output vote-fraction prediction	Clean first-gate agreement = 0.887626; Gaussian noise agreement = 0.797770	The first-gate morphology decision remained meaningful, but decreased under stronger noise.
Selective confidence	Smooth vs Featured classification	Full-test accuracy = 99.1736%; at threshold 0.88: accepted accuracy = 99.8801% with 98.4652% coverage; at threshold 0.975: accepted accuracy = 100.0000% with 95.8087% coverage	Higher confidence thresholds improved accepted-prediction reliability while reducing coverage.
Confidence and error behaviour	Galaxy Zoo 37-output prediction diagnostics	Mean confidence: correct = 0.750, incorrect = 0.578; mean sample RMSE: correct = 0.065, incorrect = 0.096	Incorrect predictions were associated with lower confidence and higher sample-level error.
Visual explanation evidence	Saliency-based explanation examples	Saliency maps visually concentrated around galaxy regions in selected lowest-error and highest-error examples	The visual evidence supports that the model attention was generally located on relevant galaxy structures.

As can be seen from the quantitative evidence in Table 9, the framework was evaluated not just in terms of accuracy and RMSE but also in terms of robustness, reliability, and visual explanations. Robustness assessments showed the degradation under the stronger perturbations, while selective-confidence analysis showed the increase of reliability by accepting only higher-confidence cases. Confidence/error diagnostics showed that incorrect predictions were associated with lower confidence and higher sample-level RMSE. Thus, all these results proved reliability-oriented design of the proposed framework.

Robustness Evidence. Robustness assessments were carried out to check the stability of the framework when the input galaxy images were subjected to controlled visual perturbations, including changes in brightness, contrast, blur, and Gaussian noise.

Robustness Evidence under Controlled Perturbations

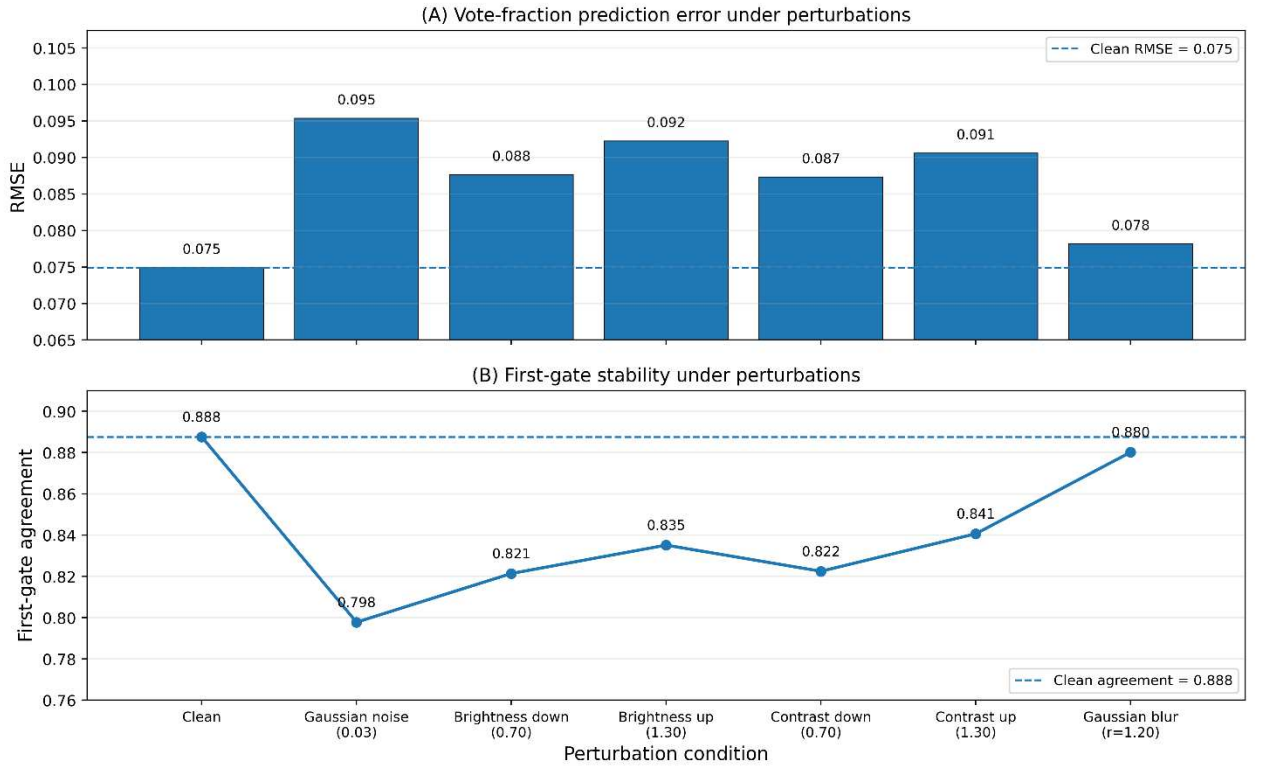


Figure 25. Framework-level robustness evidence under controlled image perturbations.

The results showed that moderate perturbations had limited effect on the model performance, while stronger Gaussian noise caused more pronounced degradation. It implied that the framework was stable under common image-quality changes but more sensitive to perturbations which hid the faint morphological structures.

Confidence and Uncertainty Evidence. Confidence and uncertainty analysis was needed to check whether the framework could distinguish between more reliable and less reliable predictions as some of the galaxy images could contain ambiguous or weak morphological features.

Confidence and Error Evidence for Correct and Incorrect Predictions

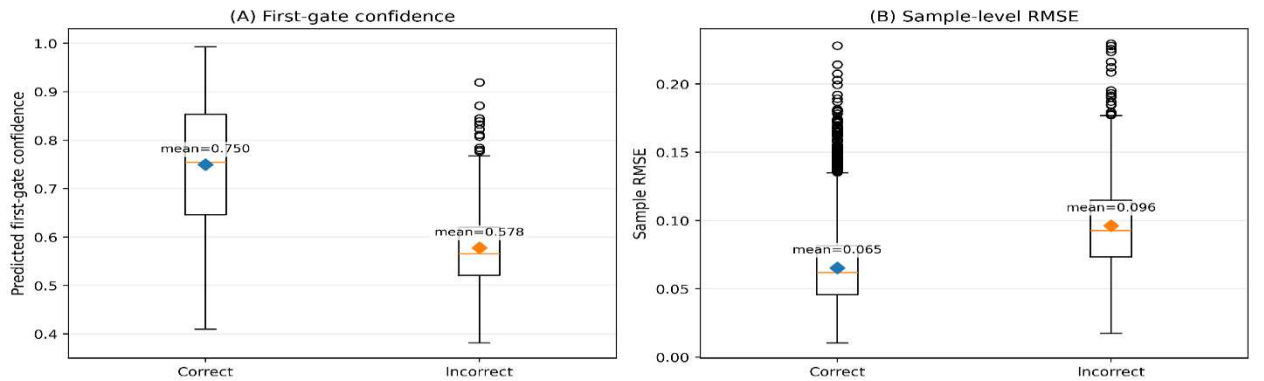


Figure 26. Confidence and error evidence for correct and incorrect galaxy morphology predictions.

The uncertainty analysis showed that the incorrect predictions were associated with weaker certainty compared to the correct predictions. Besides, selective-confidence results showed that the prediction reliability could be increased by accepting only higher-confidence cases and reporting the corresponding coverage. Thus, these results proved the usefulness of confidence-aware decision handling for the framework.

Visual Explanation Evidence. Visual explanations were needed to find out which image regions influenced selected model predictions and thus provide interpretation by revealing whether the attention of the model was concentrated on meaningful galaxy regions or irrelevant background regions.

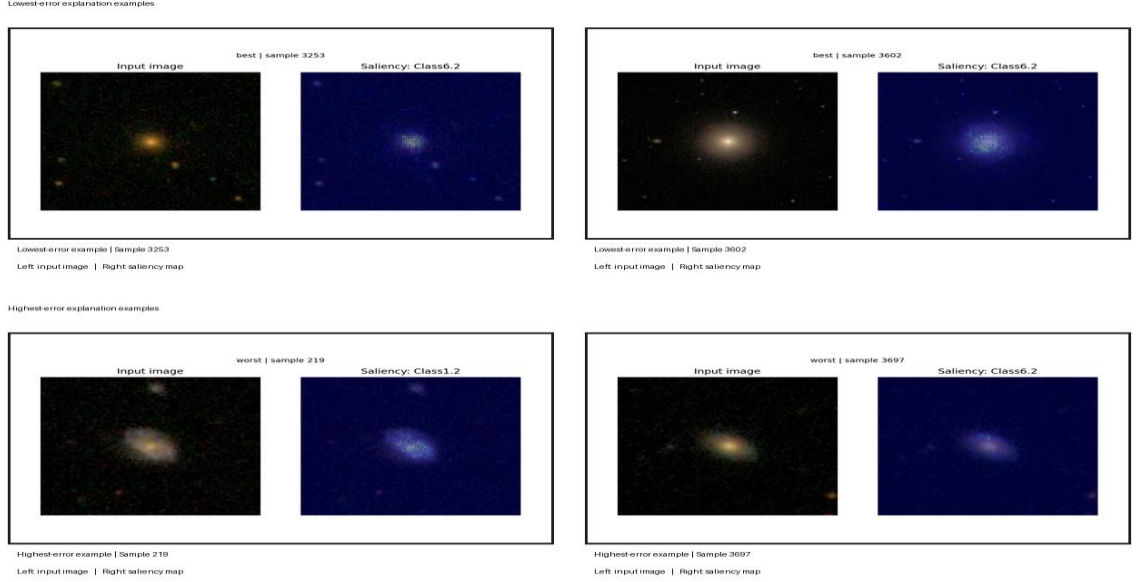


Figure 27. Saliency-based visual explanation evidence for selected lowest-error and highest-error Galaxy Zoo predictions.

Thus, saliency evidence usually concentrated on meaningful galaxy regions, such as central brightening, disk-like structure, or elongated morphology.

Overall, the reliability-oriented evaluation showed that the proposed framework was assessed not only in terms of accuracy and RMSE. Robustness results proved the stability under moderate perturbations but sensitivity to stronger perturbations. Confidence and uncertainty evidence showed that lower-reliability predictions could be identified and managed more carefully. Visual explanation evidence helped to interpret model decisions. All these results increased the usefulness of the framework as they integrated performance, stability, confidence, and interpretability into one evaluation.

4.6. Discussion

The following section highlights the key results from Section 4 and places the proposed architecture in the scope of prior research on galaxy morphology classification. Given that the experiments used two different directions for evaluation, the comparative analysis is divided into two sections: Galaxy10 DECaLS hard-label classification and Galaxy Zoo-related vote fraction evidence.

Table 10. Comparison with Galaxy10 DECaLS hard-label classification studies.

Study / Evidence	Dataset / Task	Method / Model Family	Reported Result	Position Relative to This Framework
Luo et al. [8]	Galaxy10 DECaLS classification	Semi-supervised GAN-based learning	~84% accuracy	Relevant Galaxy10 reference under limited-label learning.
Yao et al. [9]	Galaxy10 DECaLS classification	DCGAN-based data augmentation	Up to ~95% accuracy	Strong augmentation-based Galaxy10 classification reference.
Wei et al. [16]	Galaxy10 DECaLS classification	Vision Transformer	~94% accuracy	Strong transformer-based Galaxy10 comparison point.
Zhang et al. [18]	Galaxy10 DECaLS classification	CNN with attention mechanism	~96% accuracy	Strong attention-based Galaxy10 classification benchmark.
This study: Galaxy10 DECaLS hard-label task	Galaxy10 DECaLS ten-class classification	ConvNeXtV2 transfer-learning CNN model	91.58% accuracy; macro F1 = 0.9073	Provides the main hard-label classification result and connects it with class-level error analysis, robustness, confidence behavior, and visual explanation evidence.

Table 11. Comparison with Galaxy Zoo and Kaggle benchmark evidence.

Study / Evidence	Dataset / Task	Method / Evaluation Type	Reported Result	Position Relative to This Framework
Dieleman et al. [2]	Galaxy Zoo morphology classification	Rotation-invariant CNN	~91% accuracy	Provides an early CNN-based Galaxy Zoo morphology classification benchmark.
Walmsley et al. [1]	Galaxy Zoo / DECaLS morphology responses	Probabilistic deep learning / Zoobot	Vote-fraction response prediction	Provides an important reference for treating galaxy morphology as response distributions rather than only hard labels.
Pearson et al. [20]	Galaxy Zoo high-confidence morphology questions	Zoobot-based active learning and volunteer-label evaluation	Very high Smooth or Featured performance on high-confidence volunteer labels	Provides a directly related comparison point for the high-confidence Smooth vs Featured routing stage.

Ye et al. [21]	Galaxy Zoo morphology questions under cross-survey shift	Unsupervised domain adaptation for Galaxy Zoo morphology prediction	Smooth or Featured reported with strong performance under domain-adaptation evaluation	Supports the importance of robustness and dataset-shift evaluation for first-level morphology prediction.
Original Galaxy Zoo Challenge winning solution [22]	Galaxy Zoo Challenge 37-output vote-fraction prediction	Kaggle private leaderboard benchmark	Private RMSE = 0.07492	Provides the original Kaggle benchmark reference for Galaxy Zoo vote-fraction prediction.
This study: Smooth vs Featured stage	Galaxy Zoo high-confidence binary morphology routing	YOLO11x classification family	<u>99.1736% accuracy;</u> macro F1 = 0.9881	Shows strong first-level morphology routing using high-confidence vote-fraction labels.
This study: Galaxy Zoo vote-fraction task	Galaxy Zoo 37-output vote-fraction prediction	ConvNeXtV2 Large held-out test evaluation	RMSE = 0.074912; MAE = 0.042315	Extends the framework beyond single-label classification toward continuous morphology estimation.
This study: Kaggle post-deadline evidence	Galaxy Zoo Challenge 37-output vote-fraction prediction	External Kaggle scoring evidence	Public RMSE = 0.07433; Private <u>RMSE = 0.07440</u>	Provides supporting external benchmark evidence and is not treated as a final leaderboard claim.

Note: For RMSE-based scoring, lower values indicate better performance.

According to Table 10, the highest accuracy of Galaxy10 DECaLS hard-label classification from the chosen related research is about 96%, obtained by Zhang et al. [18] using convolutional neural network with an attention mechanism. In this study, Galaxy10 DECaLS hard-label branch has scored 91.58% accuracy and 0.9073 macro F1-score using ConvNeXtV2 transfer-learning CNN model. While it is not the highest Galaxy10 DECaLS accuracy ever, this constitutes the main Galaxy10 DECaLS hard-label classification result of the proposed framework and is confirmed by class-level error analysis, robustness testing, confidence behavior, and visual explanation evidence.

Table 11 shows the Galaxy Zoo-related comparison. The highest accuracy obtained in this work was achieved by the high-confidence Smooth vs Featured routing and amounted to 99.1736% accuracy and 0.9881 macro F1-score. As for Galaxy Zoo 37-output vote fraction prediction problem, the held-out test evaluation gave RMSE 0.074912 and MAE 0.042315. Also, the post-deadline evidence from Kaggle competition reached Public RMSE 0.07433 and Private RMSE 0.07440, which is better than the original Galaxy Zoo Challenge winning benchmark Private RMSE 0.07492. Note that RMSE is an error metric and the lower it is, the better performance.

Taken together, these comparisons show that the proposed framework does not depend on one single classification metric. Instead, it combines Galaxy10 DECaLS hard-label classification, Galaxy Zoo high-confidence Smooth vs Featured routing, Galaxy Zoo vote fraction prediction, and reliability-related evaluation. Thus, the scope of the framework is much wider than the scope of works that give only one single classification result or one single benchmark metric.

Despite the impressive results, their interpretation must take into account the used datasets and evaluation approaches. The problem was that some Galaxy10 DECaLS classes were still visually hard to distinguish due to overlapping morphology patterns, and some Galaxy Zoo vote fraction outputs were hard to interpret since they correspond to vote distribution rather than labels. Thus, the reliability evidence must be viewed as a complementary information that helps to interpret the results.

5. Conclusions and Future Work

5.1. Summary of the Main Findings

The main results of this study are listed according to four achievements based on the proposed framework.

Firstly, the Galaxy10 DECaLS hard-label experiment indicated that the ConvNeXtV2 transfer-learning CNN model performed well on the multi-class task achieving the held-out test accuracy score of 91.58% and the macro F1-score of 0.9073.

Secondly, the high-confidence Smooth versus Featured stage showed that reliable first-level morphology routing can be achieved using Galaxy Zoo vote-fraction labels. By applying the YOLO classification family, in particular YOLO11x, the binary classifier reached the held-out test accuracy of 99.1736% and the macro F1-score of 0.9881.

Thirdly, the Galaxy Zoo 37-output vote-fraction prediction branch showed the framework’s ability to go beyond fixed class labels and reach continuous morphology-response estimation. Applying the ConvNeXtV2 Large model, the held-out test evaluation showed the RMSE of 0.074912 and the MAE of 0.042315. The post-deadline Kaggle submission confirmed these results obtaining the Public RMSE of 0.07433 and Private RMSE of 0.07440; however, the held-out test evaluation remained the primary evaluation protocol.

Fourthly, the reliability-oriented evaluation showed that performance metrics alone were not sufficient. Robustness testing, confidence analysis, and visual explanation evidence provided additional ground for interpreting the behavior of the models under various conditions.

In sum, the above-listed findings show that the proposed framework combines hard-label classification, high-confidence morphology routing, vote-fraction prediction, and reliability-oriented evaluation into one pipeline of galaxy morphology analysis.

5.2. Achievement of Research Objectives and Answer to the Research Question

The research objectives set out in Section 1 have been accomplished through the design and evaluation of the proposed deep learning framework. The research has evaluated and compared several model families for galaxy morphology classification, investigated the impact of transfer learning, and extended the analysis from hard-label prediction to confidence-aware and vote-fraction-based morphology representation. The summary of the accomplishment of the research objectives is provided in Table 12.

Table 12. Summary of Research Objectives and Achievement Status

Research Objective	How It Was Addressed	Status
Develop deep learning models for galaxy morphology classification.	Several model families were evaluated for the Galaxy10 DECaLS task.	Achieved
Examine the role of transfer learning.	Transfer-learning CNN backbones were evaluated and compared with other model families.	Achieved
Extend the analysis beyond hard-label classification.	High-confidence classification and Galaxy Zoo vote-fraction prediction were included.	Achieved
Evaluate model performance using quantitative metrics.	Accuracy, precision, recall, F1-score, RMSE, and MAE were used according to the task type.	Achieved
Examine model reliability and interpretability.	Robustness testing, confidence analysis, and visual explanation evidence were conducted.	Achieved
Answer the main research question.	A reliability-oriented AI framework was developed and evaluated for galaxy morphology classification and analysis.	Answered within the scope of the study

The above table shows that the objectives of the study were addressed through the principal experimental components of the proposed framework. The research evaluated classification performance together with confidence-aware analysis, vote-fraction representation, robustness, and visual explanation evidence.

Based on these results, the main research question is addressed in the scope of this experimental research. The proposed framework demonstrates that the problem of galaxy morphology classification can be solved through deep learning combining the predictive performance with confidence-aware analysis, vote-fraction prediction, robustness evaluation, and visual explanation. Thus, the research question is answered by proposing the structured reliability-oriented AI framework for galaxy morphology classification and analysis.

5.3. Research Contributions and Scientific Relevance

The main contribution of this study is presenting the structured AI-based framework for galaxy morphology classification and analysis. The framework unites model-family evaluation, confidence-aware prediction, vote-fraction representation, and reliability-oriented assessment into one research direction.

From the artificial intelligence perspective, the contribution of the work lies in the evaluation of galaxy morphology classification beyond one model or a single accuracy score. Robustness testing, confidence analysis, and visual explanation evidence make the evaluation more relevant to the scientific image analysis, where the reliability and interpretability of the models are crucial.

From the astronomy perspective, the proposed framework can assist the large-scale morphology analysis by providing initial predictions, detecting high-confidence cases, and drawing attention to ambiguous examples for further investigation. Thus, the work should be regarded as the decision-support tool rather than replacing the astronomical interpretation of galaxy morphology.

5.4. Limitations

Despite the promising results, they should be considered in the scope of the datasets and experimental settings used in this research. The principal experiments used Galaxy10 DECaLS and Galaxy Zoo-related data; hence, the results cannot generalize to all astronomical surveys and imaging conditions.

Another limitation is related to the nature of galaxy morphology. Some classes are visually similar or ambiguous, especially in cases of images showing weak structures or a combination of morphological features. This is why the hard-label classification becomes difficult to perform explaining some confusion between the classes.

Besides, the robustness and visual explanation analyses should be considered as supportive evidence rather than the complete evaluation of all possible observational conditions and physical interpretations. These limitations define the scope of the conclusions and call for broader validation in future studies.

5.5. Future Work

The future work can take the following directions regarding the extension of the proposed framework. Firstly, it is possible to investigate the performance of the framework on other astronomical datasets and various sky surveys in order to assess the generalization across different image qualities, resolutions, noise levels, and observational conditions.

Another research direction is related to the improvement of uncertainty estimation and confidence calibration. While the confidence-aware analysis has been applied, future research may use more advanced probabilistic or Bayesian methods to separate the reliable predictions from the uncertain and ambiguous ones to support the human-in-the-loop analysis and flag low-confidence instances for expert interpretation.

Further research can advance the reliability and interpretability components. The robustness testing may be broadened through the additional simulations of observations or domain adaptation methods, and visual explanation approaches can be evaluated quantitatively by comparing the highlighted regions with known morphological structures like spiral arms, bars, bulges, and edge-on galaxies.

Lastly, the framework may be further developed towards the creation of the tool for the AI-assisted analysis of galaxy morphology. Such a tool will provide initial predictions, confidence indicators, vote-fractions estimation, and explanation outputs for researchers dealing with the large galaxy images datasets. These research directions will help advance the framework from the controlled experimental evaluation towards broader application in astronomical image analysis.

5.6. Conclusion

This study presents the structured deep learning framework for galaxy morphology classification and analysis. The framework combines hard-label classification, high-confidence morphology routing, Galaxy Zoo vote-fraction prediction, and reliability-oriented evaluation into one research direction.

The results show that the deep learning models can learn meaningful morphological patterns from the images and achieve good predictive performance. At the same time, the results show that the accuracy alone is not enough for scientific image analysis; hence, robustness testing, confidence analysis, and visual explanation evidence are used to support the more reliable interpretation of model behavior

Overall, this study provides a broader AI-based approach to galaxy morphology analysis combining the classification performance and confidence-aware evaluation. This makes the proposed framework the basis for further research and development of more reliable AI-assisted tools in astronomical image analysis.

References

- [1] [Walmsley, M., et al., 2022] Walmsley, M., Lintott, C., Galloway, M., et al. “Galaxy Zoo DECaLS: Detailed visual morphology measurements from volunteers and deep learning for 314,000 galaxies”. *Monthly Notices of the Royal Astronomical Society* 509(3), pp. 3966–3988, 2022. DOI:10.1093/mnras/stab2093
- [2] [Dieleman, S., et al., 2015] Dieleman, S., Willett, K. W., and Dambre, J. “Rotation-invariant convolutional neural networks for galaxy morphology prediction”. *Monthly Notices of the Royal Astronomical Society* 450(2), pp. 1441–1459, 2015. DOI:10.1093/mnras/stv632
- [3] [Bhambra, P., et al., 2022] Bhambra, P., Joachimi, B., and Lahav, O. “Explaining deep learning of galaxy morphology with saliency mapping”. *Monthly Notices of the Royal Astronomical Society* 511(4), pp. 5032–5041, 2022. DOI:10.1093/mnras/stac368
- [4] [Hubble, E., 1936] Hubble, E. “The Realm of the Nebulae”. *Yale University Press*, 1936.
- [5] [de Vaucouleurs, G., 1959] de Vaucouleurs, G. “Classification and Morphology of External Galaxies”. *Handbuch der Physik* 53, pp. 275–310, 1959. DOI:10.1007/978-3-642-45952-8_12
- [6] [Lintott, C. J., et al., 2008] Lintott, C. J., Schawinski, K., Slosar, A., et al. “Galaxy Zoo: morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey” *Monthly Notices of the Royal Astronomical Society* 389(3), pp. 1179–1189, 2008. DOI:10.1111/j.1365-2966.2008.13689.x
- [7] [Domínguez Sánchez, H., et al., 2018] Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., Tuccillo, D., and Fischer, J. L. “Improving galaxy morphologies for SDSS with Deep Learning”. *Monthly Notices of the Royal Astronomical Society* 476(3), pp. 3661–3676, 2018. DOI:10.1093/mnras/sty338
- [8] [Luo, Z., et al., 2025] Luo, Z., Chen, J., Chen, Z., et al. “Galaxy Morphology Classification via Deep Semi-supervised Learning with Limited Labeled Data”. *The Astrophysical Journal Supplement Series* 267(2), Article 45, 2025. DOI:10.3847/1538-4365/ad1f0c
- [9] [Yao, H., et al., 2024] Yao, H., Liu, Y., and Zhao, D. “A Galaxy Image Augmentation Method Based on Few-shot Learning”. *Research in Astronomy and Astrophysics* 24(3), Article 035005, 2024. DOI:10.1088/1674-4527/acf92c
- [10] [Le, T., et al., 2025] Le, T., Ibrahim, N., Nguyen, T., et al. “Enhanced Comparative Analysis of Pretrained and Custom Deep Convolutional Neural Networks for Galaxy Morphology Classification”. *Engineering Proceedings* 89(1), Article 36, 2025. DOI:10.3390/engproc2025089036
- [11] [Xie, Y., et al., 2024] Xie, Y., Zhang, H., and Wang, L. “Galaxy Morphology Classification Based on VGG19 Deep Convolutional Neural Network”. *Astrophysics and Space Science*, Article 369, 2024. DOI:10.1007/s10509-024-04124-8
- [12] [Cheng, T., 2022] Cheng, T. “Vision Transformer Applications for Galaxy Morphology Classification”. *The Astrophysical Journal* 935, Article 34, 2022. DOI:10.3847/1538-4357/ac6a52
- [13] [Cao, J., et al., 2024] Cao, J., Wu, H., Zhang, Y., et al. “Galaxy morphology classification based on Convolutional Vision Transformer”. *Astronomy & Astrophysics* 683, A44, 2024. DOI:10.1051/0004-6361/202348544
- [14] [Urechiatu, R., et al., 2024] Urechiatu, R., and Frincu, M. “Improved Galaxy Morphology Classification with Convolutional Neural Networks”. *Universe* 10(6), Article 230, 2024. DOI:10.3390/universe10060230
- [15] [Bom, C. R., et al., 2024] Bom, C. R., Mendes de Oliveira, C., Rembold, S. B., et al. “An extended catalogue of galaxy morphology using deep learning in the S-PLUS survey”. *Monthly Notices of the Royal Astronomical Society* 528(3), pp. 4188–4203, 2024. DOI:10.1093/mnras/stae277
- [16] [Wei, Y., et al., 2024] Wei, Y., Zhang, H., and Liu, Q. “Galaxy morphology classification using Vision Transformer architectures”. *Astronomy and Computing* 48, Article 100827, 2024. DOI:10.1016/j.ascom.2024.100827

- [17] [Almeida, J., et al., 2024] Almeida, J., Rodrigues, F., and Silva, P. “Transfer Learning for Galaxy Morphology Classification using Deep Convolutional Networks”. *Astrophysics and Space Science* 369, pp. 101–110, 2024. DOI:10.1007/s10509-024-04175-x
- [18] [Zhang, Y., et al., 2024] Zhang, Y., Li, H., Chen, X., and Wang, J. “Attention-based Deep Learning Framework for Galaxy Morphology Classification”. *Astronomy and Computing* 47, Article 100810, 2024. DOI: 10.1016/j.ascom.2024.100810
- [19] [Zhu, X.-P., et al., 2019] Zhu, X.-P., Dai, J.-M., Bian, C.-J., Chen, Y., Chen, S., and Hu, C. “Galaxy morphology classification with deep convolutional neural networks”. *Astrophysics and Space Science* 364, Article 55, 2019. DOI:10.1007/s10509-019-3540-1
- [20] [Pearson, J., et al., 2026] Pearson, J., Dickinson, H., Serjeant, S., Walmsley, M., Fortson, L., Kruk, S., Masters, K. L., Simmons, B. D., Smethurst, R. J., Lintott, C., et al. “Galaxy Zoo: Cosmic Dawn–morphological classifications for over 41,000 galaxies in the Euclid Deep Field North from the Hawaii Two-0 Cosmic Dawn survey”. *Monthly Notices of the Royal Astronomical Society* 546(3), Article staf2250, 2026. DOI:10.1093/mnras/staf2250
- [21] [Ye, R., et al., 2025] Ye, R., Shen, S., de Souza, R. S., Xu, Q., Chen, M., Chen, Z., Ishida, E. E. O., Krone-Martins, A., and Durgesh, R. “From Galaxy Zoo DECaLS to BASS/MzLS: detailed galaxy morphology classification with unsupervised domain adaptation”. *Monthly Notices of the Royal Astronomical Society* 537(2), pp. 640–649, 2025. DOI:10.1093/mnras/staf008
- [22] [Dieleman, S., 2014] Dieleman, S. “My solution for the Galaxy Zoo challenge”. *Sander Dieleman Blog*, 2014. Available online: <https://sander.ai/2014/04/05/galaxy-zoo.html>