



Deep Learning Framework for Galaxy Morphology Classification and Analysis

Mohammed Essa - Ahmad Yousef Alkahlout - Mohammed Taleb Mohammed
Master of Science in Artificial Intelligence - College of Informatics - Midocean University -
Union of Comoros

253000753@st.midocean.edu.km, 252000097@st.midocean.edu.km

Abstract:

Classification of galaxy morphology is important for studying the formation and evolution of galaxies. With the rapid growth of imaging datasets from surveys such as SDSS and DECaLS, manual classification has become increasingly challenging in large-scale studies. This study presents a confidence-aware deep learning framework that combines hard-label morphological classification with continuous prediction of Galaxy Zoo vote fractions. The framework was evaluated across different model families, including convolutional neural networks, transfer-learning models, transformer-based models, and YOLO classifiers. Galaxy10 DECaLS was used for ten-class morphological classification, while the Galaxy Zoo Challenge dataset was used for continuous prediction of 37 vote-fraction outputs. Accuracy, Recall, and Macro F1-score were used to evaluate the hard-label tasks, whereas RMSE and MAE were used for continuous prediction. Model reliability was examined through robustness testing, confidence analysis, Grad-CAM, and saliency maps. The ConvNeXtV2 transfer-learning model achieved a test accuracy of 91.58% on the ten-class classification task, while the YOLO11x model achieved 99.17% accuracy on the high-confidence Smooth-versus-Featured classification task. For the 37-output Galaxy Zoo prediction task, the selected model achieved an RMSE of 0.0749. These results demonstrate that the proposed framework supports accurate, interpretable, and confidence-aware galaxy morphology analysis.

Keywords: galaxy morphology; deep learning; Galaxy Zoo; Galaxy10 DECaLS; confidence-aware classification.

1. Introduction

Classification of galaxy morphology stays one of the essential issues in the field of astronomy because observable structures of galaxies contain valuable information about origins and evolutionary processes in them. Due to the development of astronomical science, parameters like shape, luminosity, spiral structures, and central bulges are now valuable indicators of physical properties and galaxy formation history. Therefore, the issue of galaxy morphology classification stays one of the relevant research directions (Walmsley et al., 2022).



Figure 1. An example of galaxy morphology diversity in the context of automated galaxy classification. Source: NASA/Hubble.

As it can be seen in Figure 1, galaxies have quite a variety of images. While some of them have clear morphology, others have a more complex shape and unclear boundaries of morphological type. Previously, galaxy classification was done manually by astronomers (Walmsley et al., 2022). Even though this method helped in making some progress, nowadays there is a large number of galaxy images obtained by large sky surveys (SDSS and DECaLS), which makes manual galaxy classification difficult and inefficient (Walmsley et al., 2022).

In recent decades, the use of machine learning has increased significantly. Convolutional Neural Networks that can be trained on images show the level of performance close to humans (Dieleman et al., 2015). In addition to accuracy, more metrics have gained importance in terms of scientific applications: the model has to use proper features, be robust enough to different conditions of image acquisition and provide predictions under uncertainty (Bhambra et al., 2022).

This paper investigates galaxy morphology classification using deep learning with a focus on predictability, robustness, uncertainty and interpretability (Bhambra et al., 2022). For galaxy morphology classification, the Galaxy10 DECaLS dataset will be used as main source of data because it has labeled samples easily accessible. Even though some previous works focused on the separate aspects of galaxy classification (like semi-supervised learning or transformers), in this work the choice of modern architecture and data pre-processing was made in order to create a feasible framework (Luo et al., 2025). Thus, the usage of deep learning in the field is not only about getting accurate results but also having models stable under changing conditions and providing more insight into the decision-making process. Hence, galaxy morphology classification is framed as a problem of performance, stability and interpretability (Cao et al., 2024).

1.1. Problem Statement

Even though many deep learning papers deal with galaxy morphology classification and mostly focus on accuracy and F1-score metrics, those metrics still cannot cover all aspects of scientific analysis. First of them is interpretability because if we classify the galaxy as spiral, we need to make sure that it relies on spiral features and the central bulge and not something else in the image. Furthermore, the variation of conditions in which galaxy image was captured, such as illumination, noise, and contrast changes, can decrease the performance if the images are perturbed. Reliable and robust predictions are needed in scientific application, even for altered images.

Besides, labeling each galaxy with one definite label does not reflect the uncertainty connected with some morphologies. Galaxies may have mixed morphology, and in this case, confidence-based representation would provide more information than exclusive labels. Although the existing deep learning solutions can give us good predictions, no universal framework that covers interpretability, robustness, and prediction uncertainty is provided. For evaluation of the proposed solution, Galaxy10 DECaLS dataset will be used as a primary source of labeled images appropriate for deep learning experiments.

1.2. Main Contribution

In order to overcome those issues, the current study suggests a new, reliable framework that extends automated galaxy morphology classification with the structured evaluation of deep learning architectures (baseline CNN, transfer learning like ResNet and EfficientNet) and performance metrics. The main contribution consists in the creation of confidence-aware framework that moves from a controlled hard-labeling approach to the continuous confidence-based one with Galaxy Zoo vote-fractions annotations, thus making it possible to represent uncertain and transitional features. Moreover, this research introduces a structured robustness evaluation approach that will allow evaluating model stability under controlled image perturbations and observational differences. Finally, by introducing explainable AI techniques (Grad-CAM, saliency map, etc.), this work estimates prediction uncertainty and make sure that models use only scientifically relevant galactic structures instead of noise, thus combining high accuracy with scientific reliability.

2. Previous Studies

The following section describes briefly background information and previous work related to the topic of galaxy morphology classification using deep learning approaches.

Walmsley et al. (2022) 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.

Bhambra et al. (2022) 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.

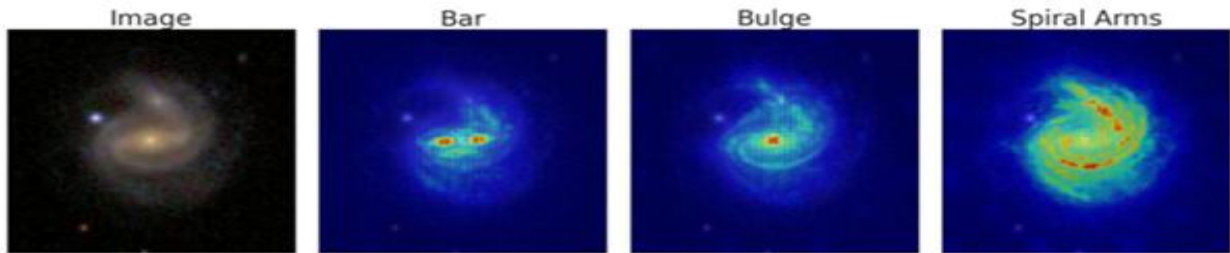


Figure 2. Example of a saliency map highlighting discriminative regions in galaxy morphology classification.
Source: Adapted from Bhambra et al. (2022).

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.

Luo et al. (2025) 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. (2024) 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. (2025) 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. (2024) 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 researchers 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 (2022) 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. (2024) 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.

Wei et al. (2024) 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.

Zhang et al. (2024) 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 studies. 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.

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

Table 1. Comparison of related studies in galaxy morphology classification

Author	Year	Technique Used	Accuracy	Dataset	Comparison with This Research
Walmsley et al.	2022	Bayesian Deep Learning (Zoobot)	Not explicitly reported	Galaxy Zoo / DECaLS	Reference for uncertainty-aware galaxy classification
Bhambra et al.	2022	Grad-CAM & Saliency Mapping	Not focus of study	Galaxy Zoo / DECaLS	Reference study for explainable AI evaluation
Luo et al.	2025	Semi-supervised Learning (GAN-based)	~84%	Galaxy10 DECaLS	Semi-supervised learning reference under limited labels
Yao et al.	2024	DCGAN Data Augmentation	Up to ~95%	Galaxy10 DECaLS	Data augmentation reference for improving model generalization
Le et al.	2025	EfficientNetB5 Transfer Learning	~87%	Galaxy10 DECaLS	Transfer learning architecture benchmark
Xie et al.	2024	VGG19 Transfer Learning	~84.5%	Galaxy10 DECaLS	CNN transfer learning baseline comparison
Cheng	2022	Vision Transformer	~86%	Galaxy10 DECaLS	Transformer-based architecture comparison
Cao et al.	2024	Convolutional Vision Transformer (CvT)	~98%	Galaxy Zoo CANDELS	Hybrid CNN–Transformer performance benchmark
Wei et al.	2024	Vision Transformer (ViT)	~94%	Galaxy10 DECaLS	Transformer-based architecture comparison

Although recent studies have reported promising results, 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, and heavy dependence on massive, perfectly labeled datasets remains a challenge for future large-scale surveys. Crucially, a distinct gap persists in current literature where classification accuracy, uncertainty, robustness, and interpretability are treated as entirely separate paths of research. Furthermore, a clear methodological divide remains between rigid hard-label classification and the richer representation of vote fractions. The framework proposed in this study directly bridges the gap between class label classification and a more general confidence-based approach, combining these critical dimensions into a single setting.

3. Methodology

The methodology is offered as a rigorous end-to-end engineering pipeline designed to solve specific problems inherent to hard-label classification and continuous morphology estimation. The architecture performs a structural analysis following a step-by-step process starting from data ingestion, specialized multi-branch preprocessing pipelines, and various deep backbone feature extractors before reaching continuous multi-output prediction.

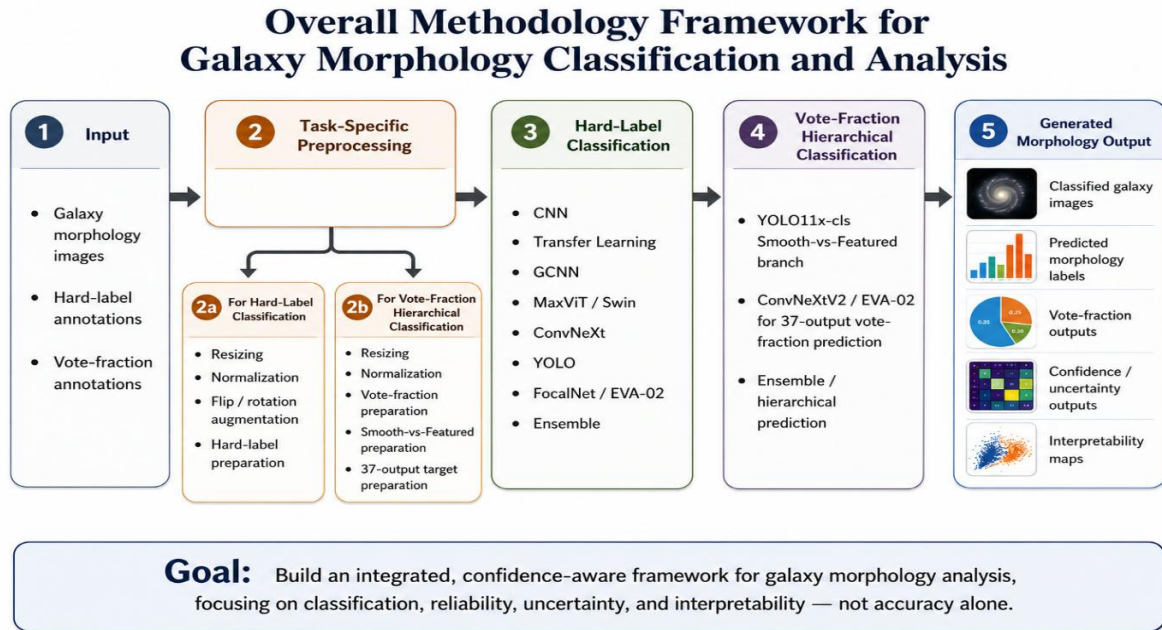


Figure 3. Overall proposed galaxy morphology methodology framework.

3.1. Input (Datasets)

To test the pipeline's capability across different classification approaches, two major datasets were utilized:

- Galaxy10 DECaLS Dataset: It offers the basis for the multi-class hard-label branch, containing discretized morphological classes across thousands of labeled images.



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

- Galaxy Zoo Challenge Dataset: To enable continuous, confidence-aware hierarchical prediction, it was used because of its ability to preserve original, uncompressed volunteer vote fraction targets.

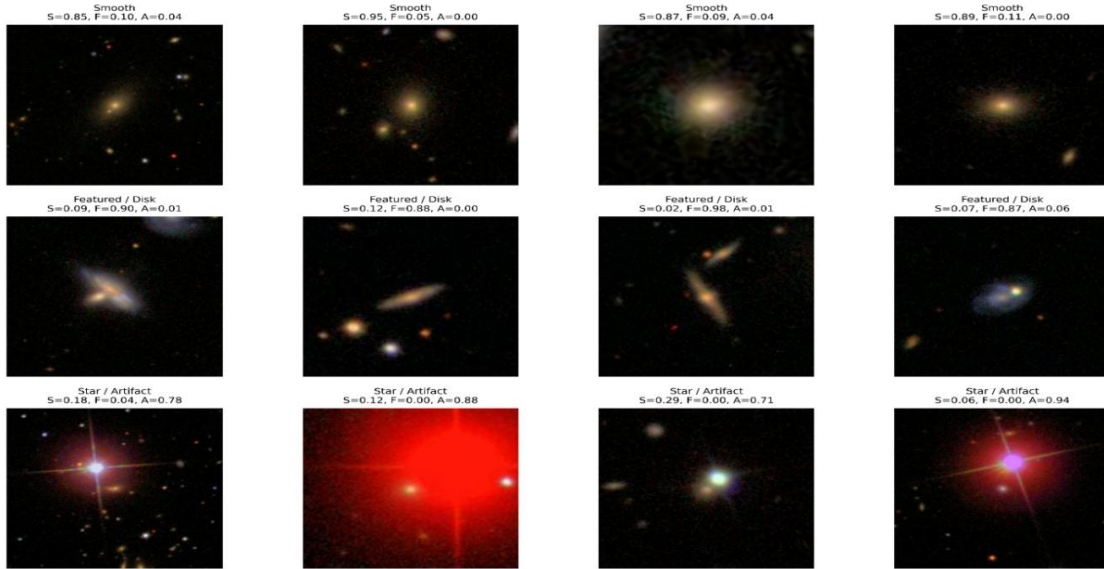


Figure 5. Sample Galaxy Zoo Challenge images with first-gate vote-fraction annotations used for hierarchical morphology classification.

3.2. Preprocessing

Since the pipeline works with both discrete classes and continuous target distributions, the preprocessing phase is deliberately separated into two parallel paths.

3.2.1. Hard-Label Preprocessing Pathway

Galaxy10 DECaLS images classified within this branch go through multi-scale spatial resizing. These images are processed with different dimensions of 64×64 , 128×128 , 160×160 , 224×224 , and 256×256 pixels to study how preserving spatial information affects feature extraction in comparison with computing costs. Channel-wise normalization according to ImageNet standard was applied to all pre-trained deep learning backbones with the following formula:

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

where $\mu = [0.485, 0.456, 0.406]$ and $\sigma = [0.229, 0.224, 0.225]$. In order to ensure orientation invariance during training while preventing information leakage, random rotations and reflections were applied to the training subset only.

3.2.2. Continuous Vote-Fraction Pathway

For the purpose of continuous prediction, the targets were preserved as a full multi-output regression vector with the whole 37-dimensional Galaxy Zoo morphology decision tree:

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

Alongside, a high-confidence binary classification branch focused on distinguishing Smooth and Featured systems was created based on isolated images when the first-gate user vote fraction is equal to or exceeds a certain confidence threshold τ :

$$\text{label} = \begin{cases} \text{Smooth}, & \text{"if Class1.1"} \geq \tau \\ \text{Featured}, & \text{"if Class1.2"} \geq \tau \end{cases} \quad (3)$$

Images meeting the above-mentioned condition were resized to a higher 448×448 pixels resolution required for training of the dedicated YOLO11x-cls architecture.

3.3. Experimental Control and Data Partitioning

To avoid any data leakage and keep validation choices and interpretability metrics clean, a strict data partitioning protocol was applied. The fully labeled Galaxy10 DECaLS dataset was separated into completely independent training, validation, and testing subsets. The numbers of images, their ratio, and roles for each subset are specified in Table 2:

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

Unlabeled imagery, such as Kaggle test data, is not included in the local training and hyperparameters tuning process at all but only used for post-deadline external validation.

3.4. Classification Models

Several distinct architectural families of deep learning models were being used in the pipeline to compare local feature extraction with global context modeling. The baseline layer blocks, dimensional transformations, and mathematical attributes implemented in these models are briefly summarized below (Walmsley et al., 2022).

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

Deep Learning CNN Baseline: Traditional convolutional neural networks represented the basic baseline, passing images through the chain of convolutional and pooling blocks to automatically detect localized structural patterns (e.g., spiral arms or galactic cores) through spatial filtering:

$$F(i, j, k) = \sum_m \sum_n \sum_c X(i + m, j + n, c) W(m, n, c, k) + b_k \quad (4)$$

Group Convolutional Networks (GCNN D16): To explicitly take the geometric nature of astronomical data into account, an orientation-aware extension is applied. This architecture embeds rotation and reflection symmetries directly into weight layers, ensuring the geometric consistency in D16 dihedral group configuration:

$$[f * \psi](g) = \sum_{x \in \mathbb{Z}^2} f(x) \psi(g^{-1}x) \quad (5)$$

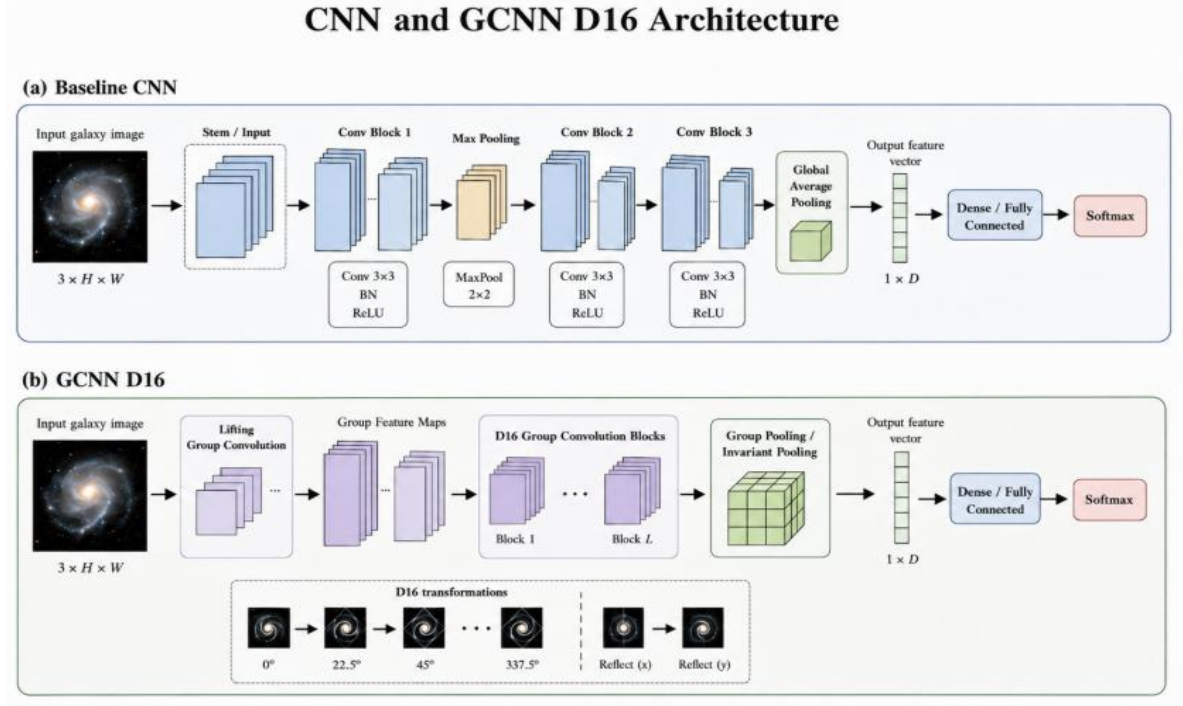


Figure 6. CNN family architecture used for local galaxy morphology feature extraction.

Transfer Learning Backbones: Advanced pre-trained models like ResNet, EfficientNet, and ConvNeXtV2 were utilized to transfer complex visual features trained on large natural images to astronomical pixel domain.

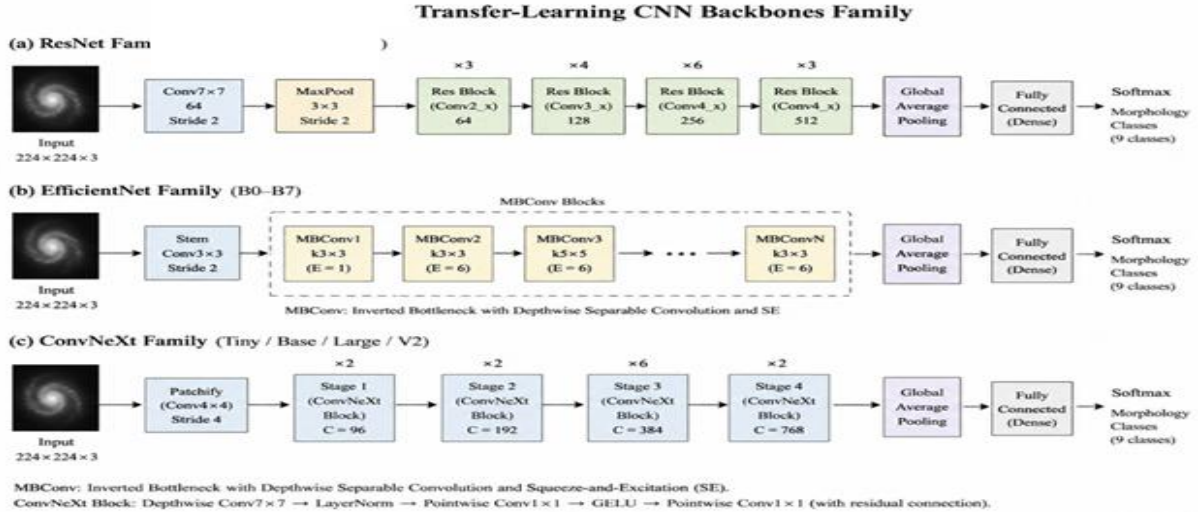
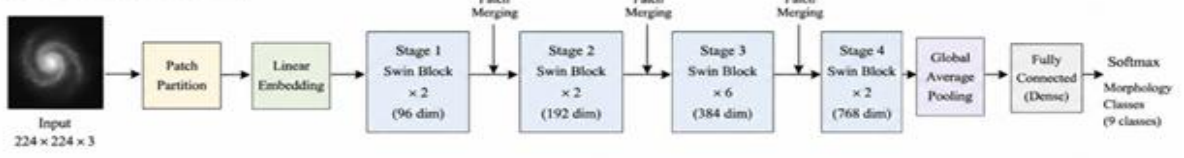


Figure 7. Transfer-learning CNN backbone family adapted for galaxy morphology classification

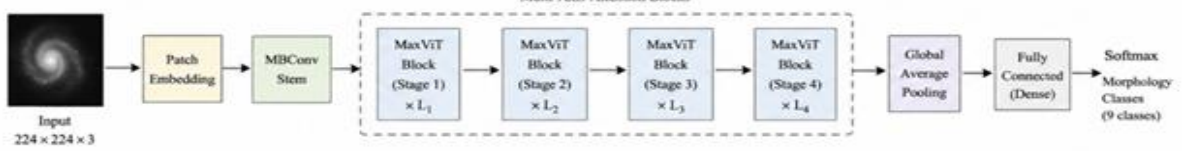
Vision Transformers and Hybrid Models: Attention mechanism blocks were added through ViT and Convolutional Vision Transformers (CvT) to handle dependencies across larger spatial layout, including outer tidal tails and low surface brightness features.

Transformer and Hybrid Attention Family

(a) Swin Transformer Family



(b) MaxViT Family



(c) FocalNet / EVA-02 Family (Hybrid Attention)

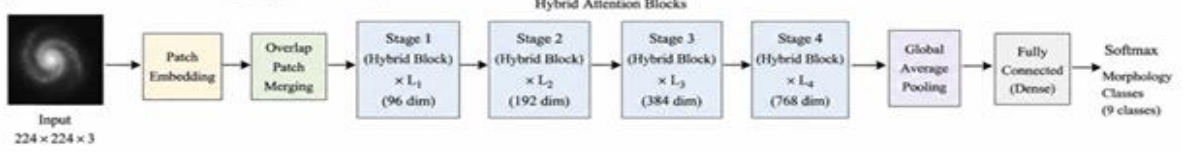


Figure 8. Transformers and hybrid attention architecture family used for broader spatial morphology reasoning.

YOLO Classification Family: For fast structural screening, a highly efficient classifier uses optimized backbone representation and routes it through a linear softmax top layer for quick image-level routing.

YOLO Classification Family

(a) YOLO Classification Model

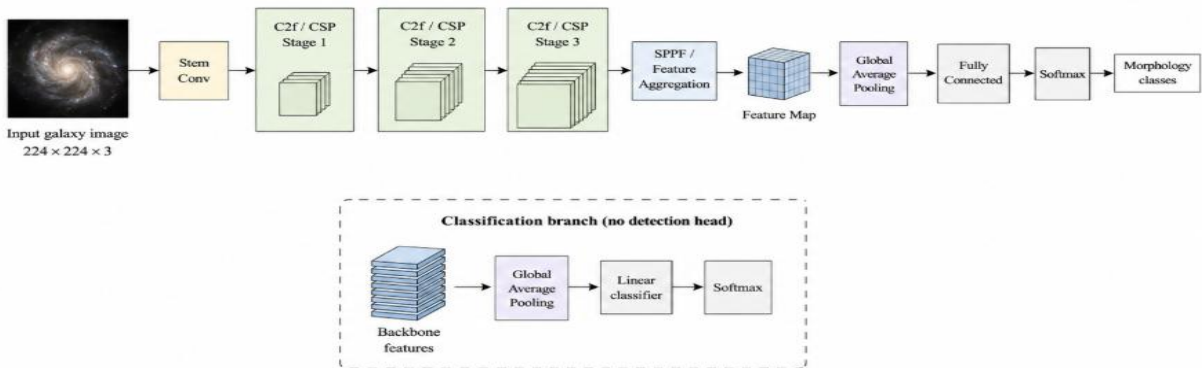


Figure .9 YOLO classification architecture family used as an efficient image-level morphology classifier

Ensemble and Hierarchical Strategy: To minimize model variance, a weighted ensemble approach was applied, aggregating predictions from different architecture families into one optimized output vector:

$$\hat{y}_j = \sum_{m=1}^M w_m \hat{y}_{j,m}, \quad \text{where} \quad \sum_{m=1}^M w_m = 1 \quad (6)$$

Ensemble and Hierarchical Classification Strategy

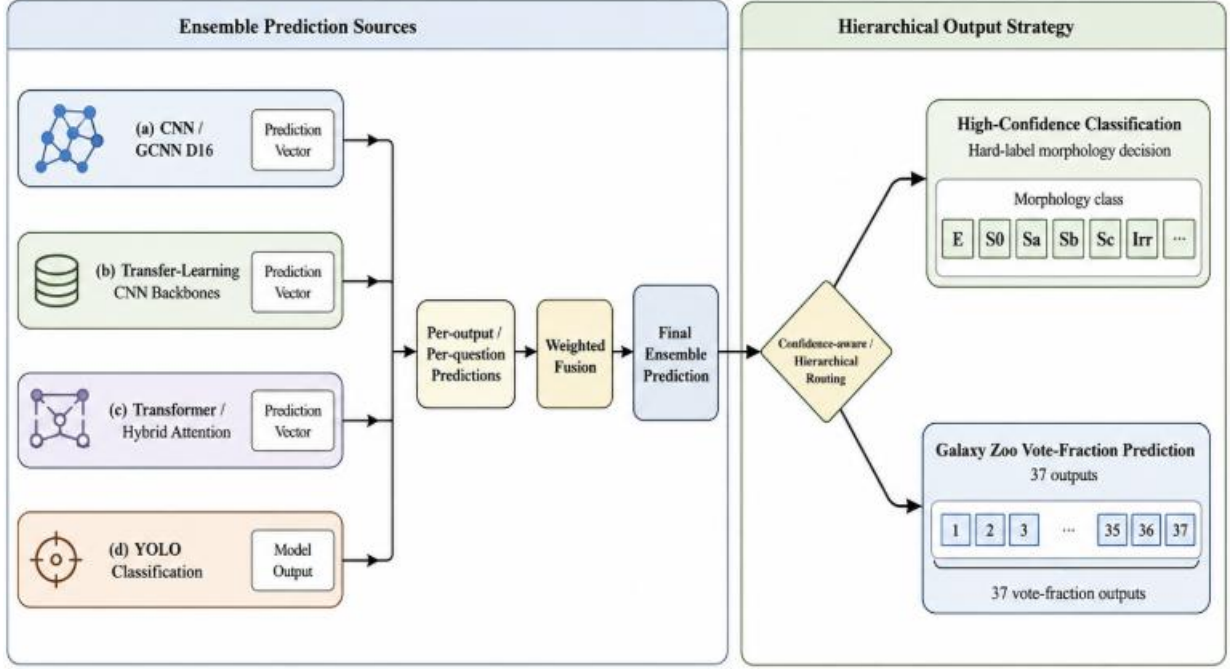


Figure 10. Ensemble and hierarchical prediction strategy connecting model families with 37-output Galaxy Zoo morphology prediction.

3.5. Training and Evaluation Protocol

Training process and evaluation metrics are adjusted according to learning goals of each particular branch:

Hard-label multi-class classification tasks were optimized with categorical cross-entropy loss function:

$$\mathfrak{J}_{\text{CE}} = - \sum_{c=1}^C y_c \log(\hat{y}_c) \quad (7)$$

Accuracy, Precision, Recall, and Macro F1-score were used as evaluation metrics.

For the continuous 37-dimensional vote-fraction tracking, the problem was considered as a multi-output regression problem, optimized with regression loss and evaluated with RMSE and MAE to accurately reflect the model's uncertainty.

Finally, the whole pipeline incorporates Explainable AI (XAI) procedures to visualize that the model focuses on true astronomical features with Grad-CAM and saliency map techniques.

4. Results and Discussion

This section provides the empirical verification of the proposed galaxy morphology framework. The performance assessment covers the aspects of discrete hard-label classification, high-confidence binary routing, continuous multi-output vote-fraction estimation, and the extensive reliability evaluation including robustness, calibration, and visual explanations.

4.1. Preprocessing Results

In order to ensure the data consistency required for the operational paths outlined in Section 3, the data ingestion pipeline was validated with regard to both hard-label and continuous targets. The output data and stratification settings are provided in Table 4.

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

Some examples of source images and the corresponding normalized spatial representations in both pipelines are depicted in Figure 11.

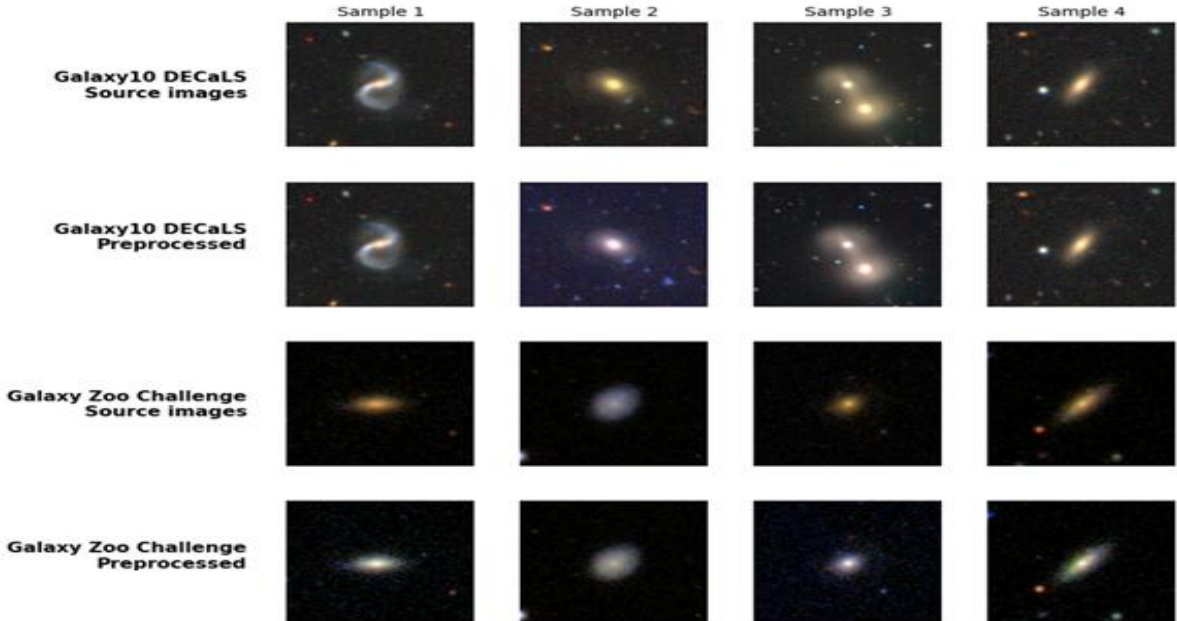


Figure 11. Representative source and preprocessed galaxy images from the Galaxy10 DECaLS and Galaxy Zoo datasets

4.2. Hard-Label Classification Results Across Model Families

Multi-class hard-label classification performance was assessed systematically for the principal families of models presented in the methodological part. Comparative experimental results on the 10-class Galaxy10 DECaLS dataset are provided in Table 5.

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 attention family	Swin Large 384 attention model	Galaxy10 full 10-class	Validation accuracy = 0.884962; Macro Recall = 0.861231; Macro F1 = 0.862763	Validation evidence	Shows the contribution of attention-based 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 presents the representative experiments selected from each model family examined in this study. The Baseline CNN and GCNN D16 experiment achieved a validation accuracy of 0.886466, a Macro Recall of 0.876281, and a Macro F1-score of 0.878177. Among the models evaluated on the full ten-class held-out test set, the ConvNeXtV2 transfer-learning CNN model achieved the best performance, with an accuracy of 0.915821, a Macro Recall of 0.908692, and a Macro F1-score of 0.907277. The Swin Large 384 attention model achieved a validation accuracy of 0.884962, a Macro Recall of 0.861231, and a Macro F1-score of 0.862763. For the YOLO11 classification family, the YOLO11x-cls and YOLO11 ensemble classifiers achieved a validation Top-1 accuracy of 0.893233, a weak-class mean recall of 0.753126, and a Macro F1-score of 0.878678. The YOLO11 and FocalNet reduced-class fusion achieved an accuracy of 0.980296, a Macro Recall of 0.971904, and a Macro F1-score of 0.970908 on the reduced seven-class validation setting. It is important to note that this final result belongs to a reduced-class diagnostic experiment and should not be directly compared with the results obtained on the full ten-class task.

Due to better test metrics on the 10-class task, the ConvNeXtV2 transfer-learning CNN model was chosen for detailed per-class analysis and uncertainty profiling.

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

The classification report of the ConvNeXtV2 architecture on the test dataset is presented below in Table 6.

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

Class	Precision	Recall	F1-score	Support
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

The test confusion matrix (Figure 12) demonstrates primarily diagonal structure. Errors are mainly localized in visually ambiguous classes: tight and loose unbarred spirals (Class 6 and Class 7) and disturbed galaxies (Class 0). Some examples of correctly and incorrectly classified images are provided in Figure 13.

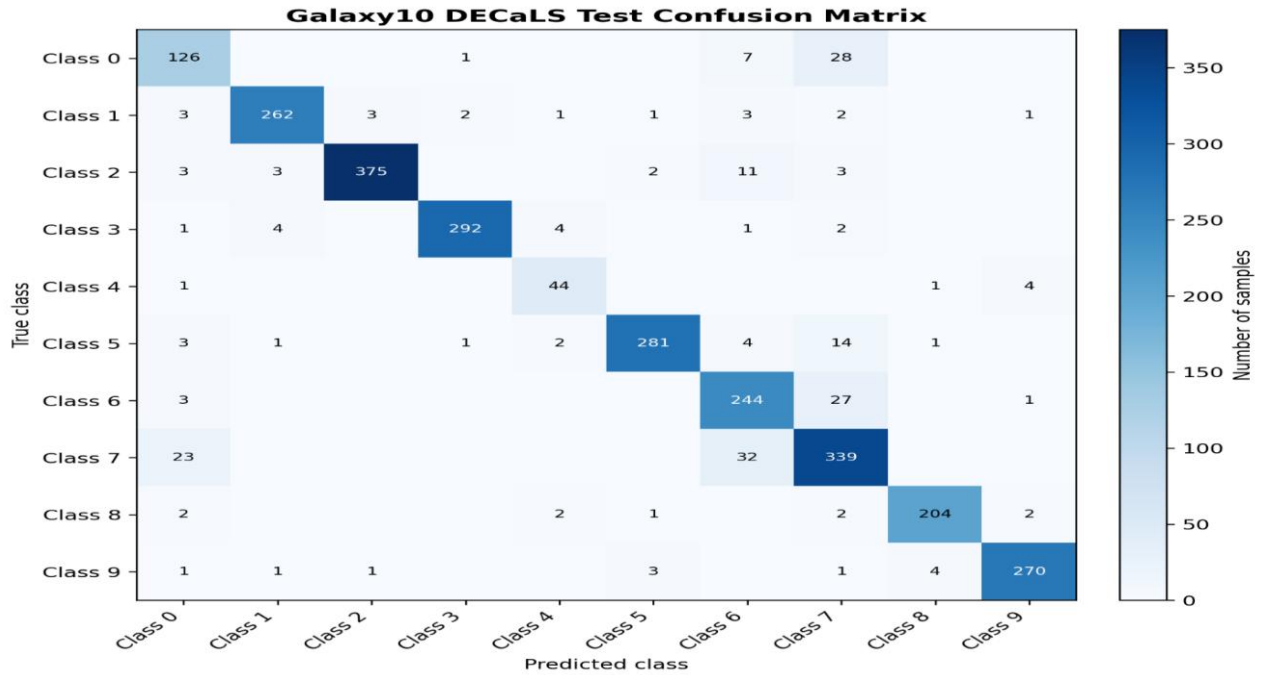


Figure 12. Galaxy10 DECaLS ten-class confusion matrix on the held-out test set.

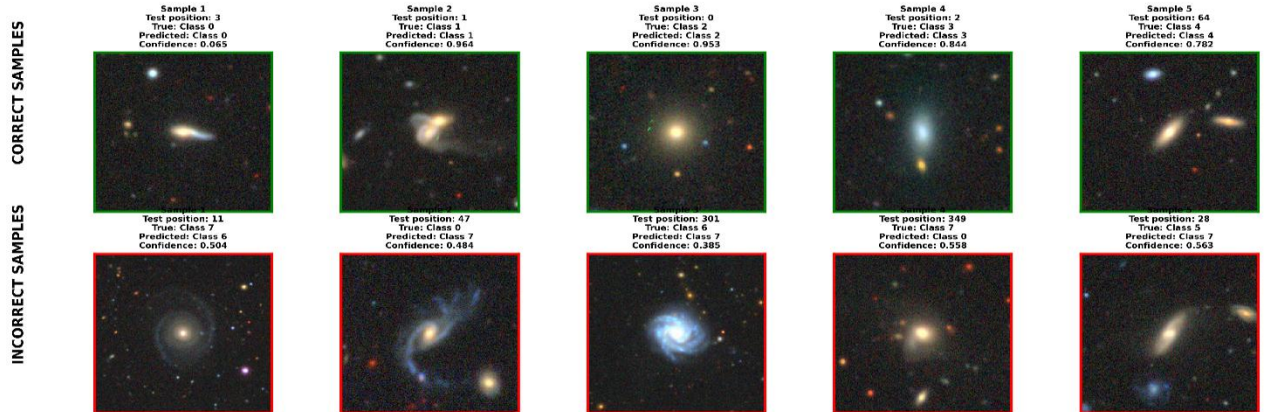


Figure 13. Representative correct and incorrect Galaxy10 DECaLS predictions from the ConvNeXtV2 transfer-learning CNN model.

In order to check if the model isolates physically significant galactic features and not just the background noise, the Grad-CAM maps were constructed as in Figure 14. The heatmaps demonstrate strong focus on scientifically relevant structures like central bulges, disks, and asymmetries.

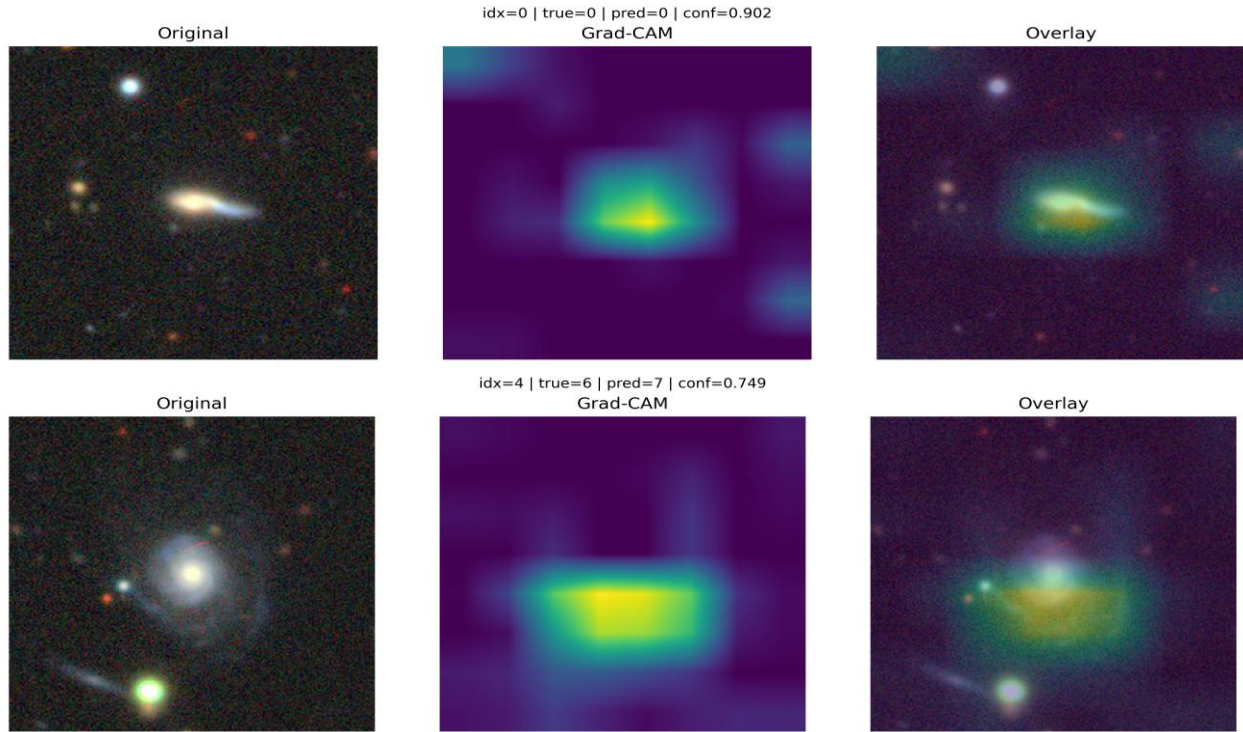


Figure 14. Grad-CAM visual explanations for selected Galaxy10 DECaLS predictions from the ConvNeXtV2 transfer-learning CNN model.

Robustness profiling under the controlled image perturbations (Figure 15) reveals that while the model is robust to moderate changes, the performance deteriorates significantly with extreme noise and blur types.

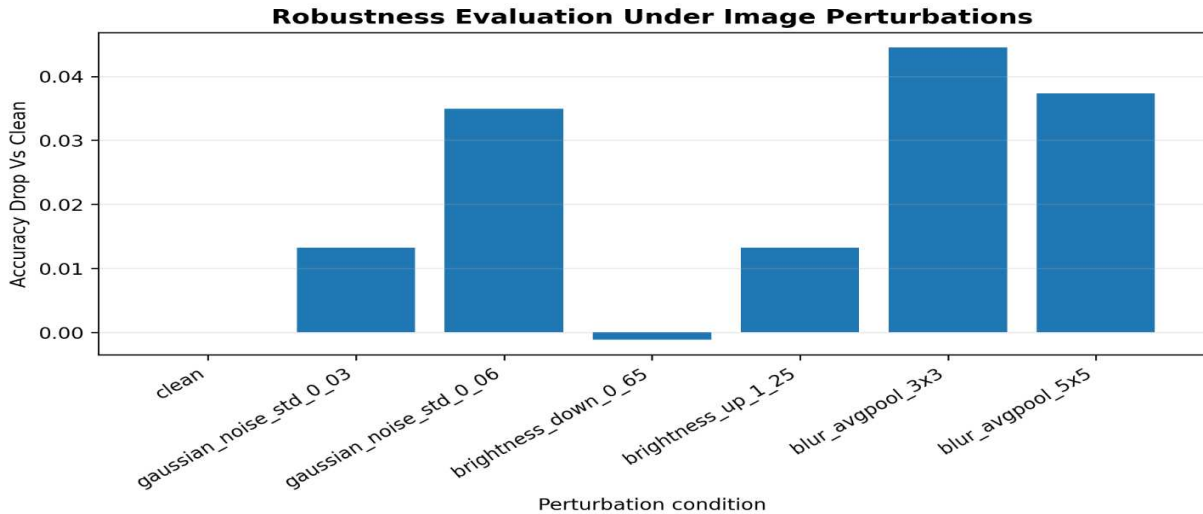


Figure 15. Robustness evaluation under controlled image perturbations.

The uncertainty distribution (Figure 16) shows that incorrect predictions correspond to more spread and lower confidence profiles than correct ones, which validates the self-awareness of the framework.

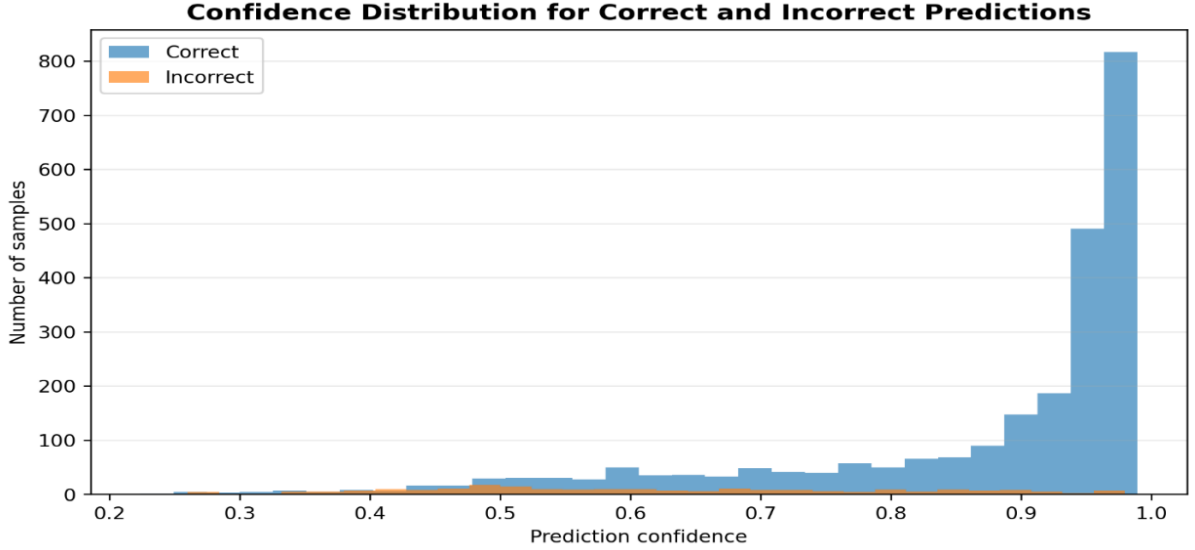


Figure 16. Uncertainty distribution for correct and incorrect Galaxy10 DECaLS predictions.

4.3. High-Confidence Smooth vs Featured Classification

The morphology routing gateway was tested through the high-confidence Smooth vs Featured binary classification path using the YOLO11x architecture. The training and evaluation datasets were filtered with a high vote fraction threshold ($\tau \geq 0.90$) in order to include only unambiguous targets (Figure 17).

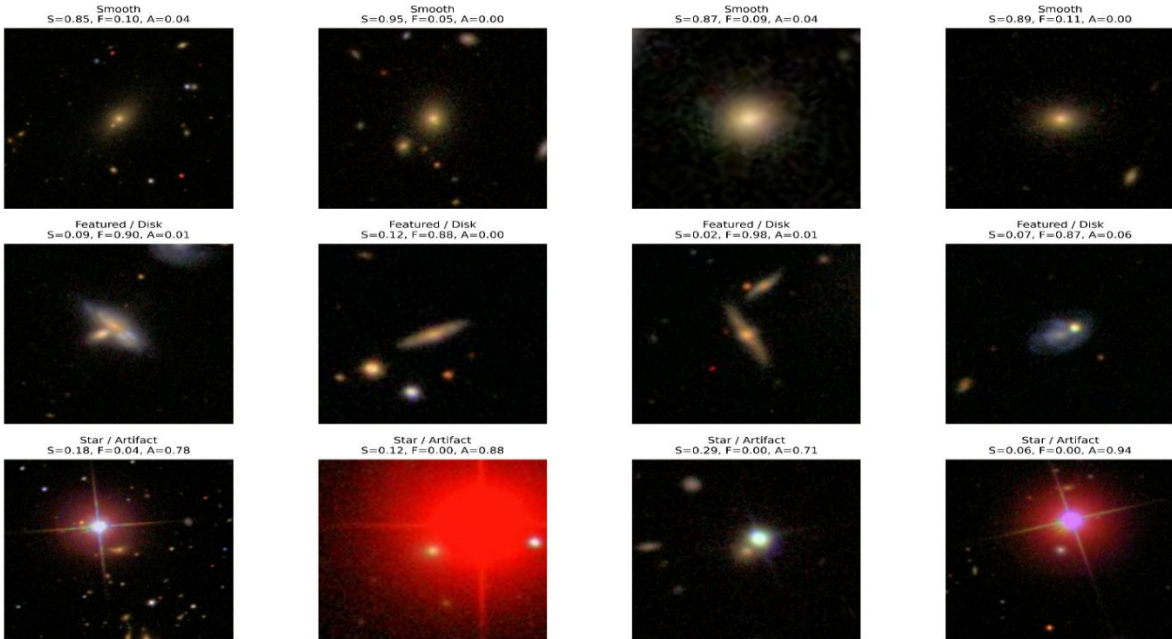


Figure 17. Galaxy Zoo first-gate vote-fraction examples used to construct the high-confidence Smooth vs Featured binary classification task.

The performance on the test dataset is summarized in Table 7 and demonstrates the accuracy 99.1736% and the macro F1-score 0.988126.

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

As seen from the binary confusion matrix (Figure 18), the errors are evenly distributed: 9 Featured galaxies were classified as Smooth and 5 Smooth ones as Featured. The qualitative error analysis (Figure 19) confirms that errors appear only in faint galaxies with low contrast.

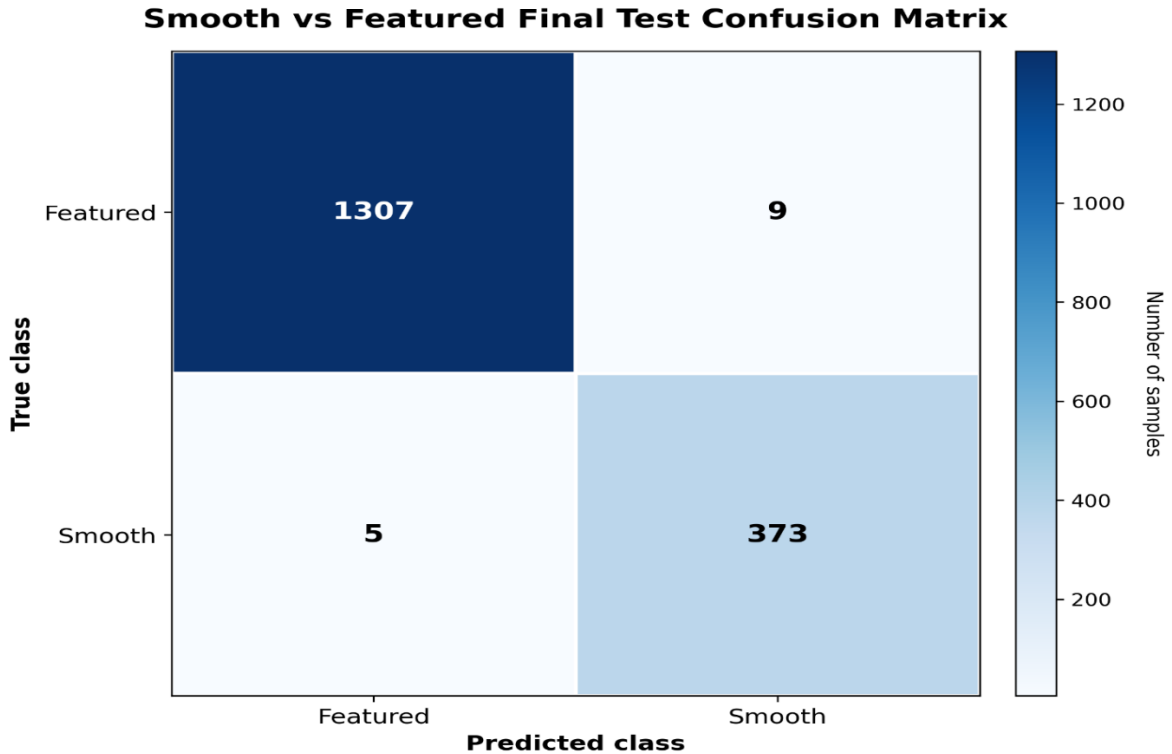


Figure 18. Official held-out test confusion matrix for the high-confidence Smooth vs Featured binary classification stage.

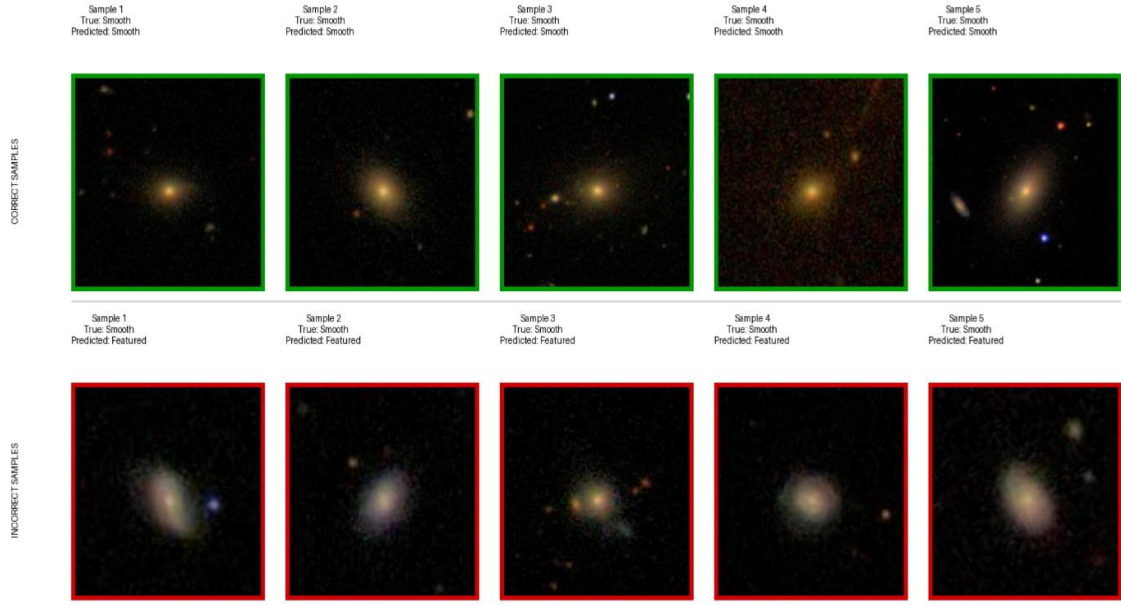


Figure 19. Representative correct and incorrect predictions from the high-confidence Smooth vs Featured classification stage.

Selective confidence filtering (Figure 20) reveals that the constraint $\tau > 0.975$ increases the routing accuracy to 100.0000% while leaving only 95.8087% of data covered. The robustness test (Figure 21) demonstrates robust performance ($\geq 99\%$) for standard perturbations, while for the high Gaussian noise the accuracy drops a bit.

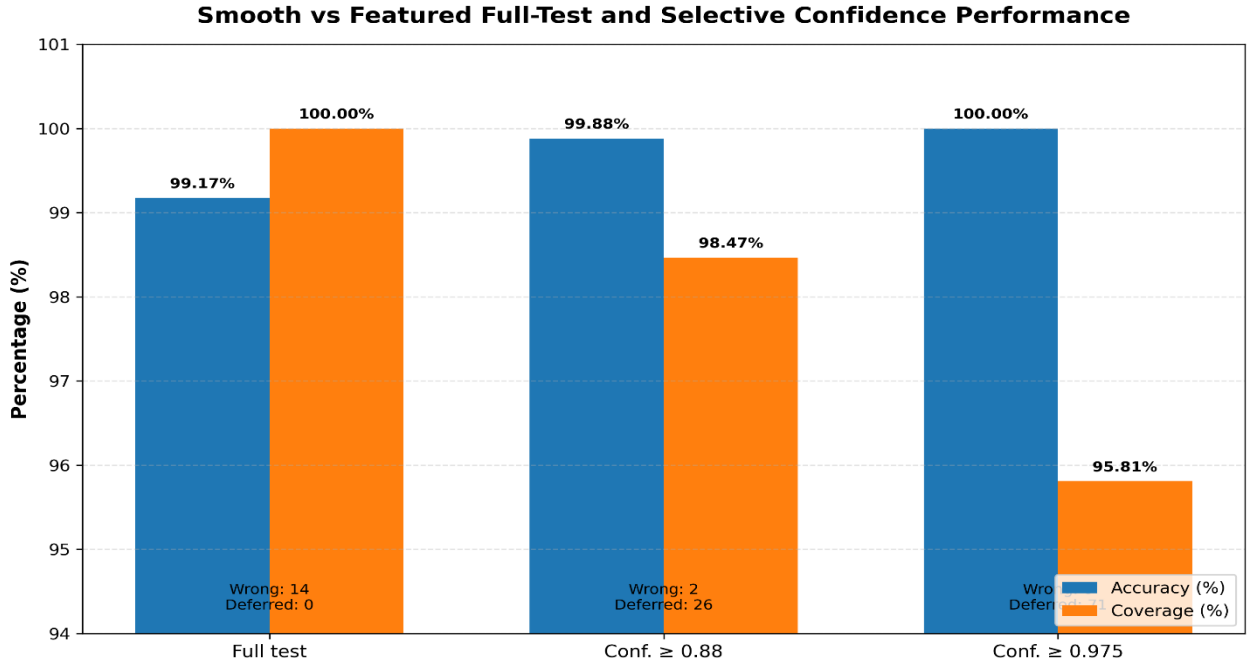


Figure 20. Full-test and selective-confidence performance of the Smooth vs Featured classification stage.

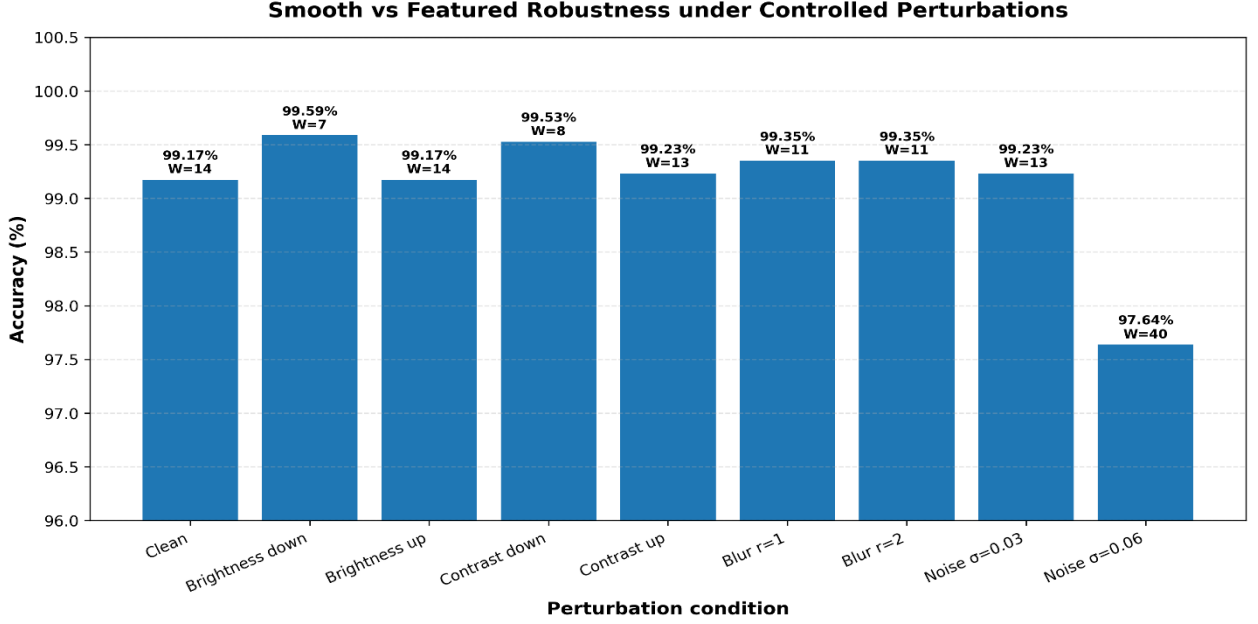


Figure 21. Robustness of the Smooth vs Featured classifier under controlled image perturbations.

4.4. Galaxy Zoo Vote-Fraction Prediction Results

To assess the morphological uncertainty beyond hard boundaries, the continuous multi-output branch was validated with the help of the ConvNeXtV2 Large architecture that predicts all 37 continuous vote fractions simultaneously. Quantitative performance results are provided in Table 8.

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

Per-output error mapping (Figure 22) indicates that error distribution depends on the complexity of the underlying questions tree, with primary features being predicted with smaller RMSE value than deeper geometric ones.

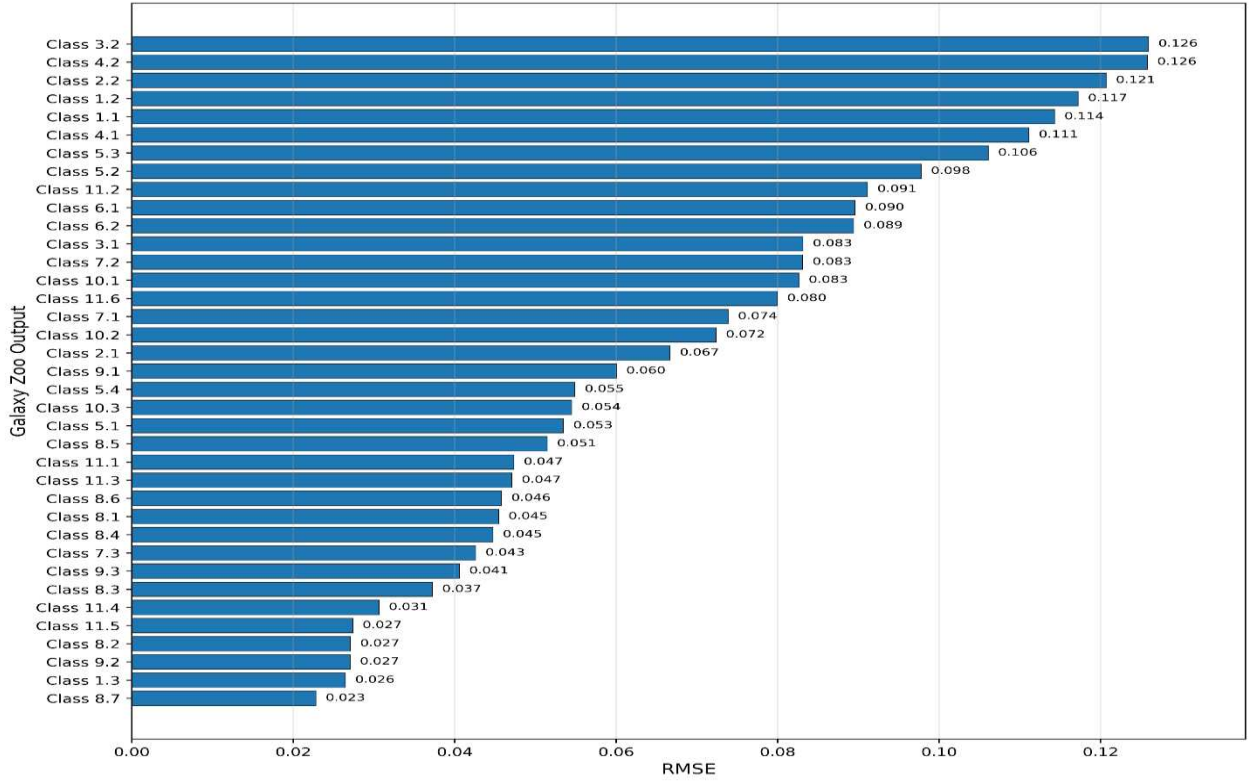


Figure 22. Per-output RMSE values for the Galaxy Zoo 37-output vote-fraction prediction task.

The first gate diagnostic filtering reveals the dominant class matching accuracy 88.7626% (Figure 23), confirming that the regression vectors preserve the structural relevance. Examples of lowest-error and highest-error targets are provided in Figure 24.

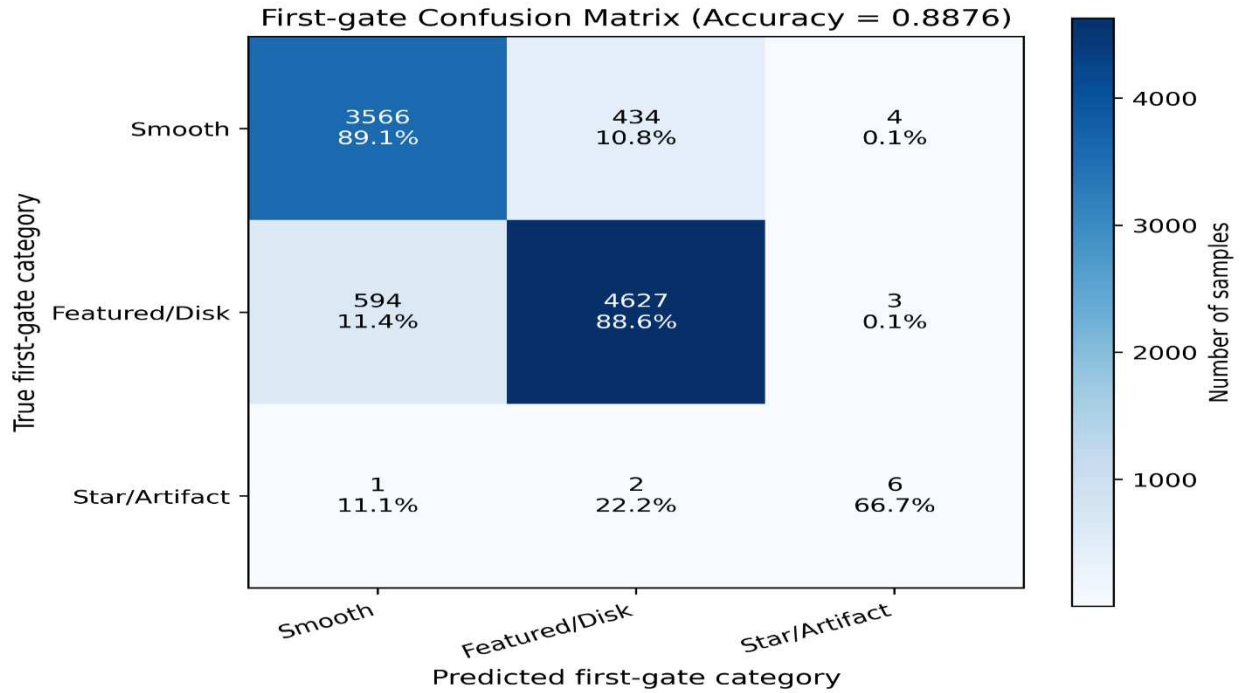


Figure 23. First-gate prediction analysis for the Galaxy Zoo 37-output vote-fraction model.



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

4.5. Reliability-Oriented Evaluation

Reliability evidence collected during the robustness testing, selective-confidence analysis, confidence/error diagnostics, and visual explanations with saliency maps is enumerated in Table 9.

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

Reliability component	Experiment / evidence source	Quantitative evidence	Interpretation
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.

Corresponding visual evidence for robustness of the pipeline (Figure 25), confidence/error diagnostics (Figure 26), and saliency maps (Figure 27) confirms that feature extraction of the model remains tightly related to physically relevant structures.

Robustness Evidence under Controlled Perturbations

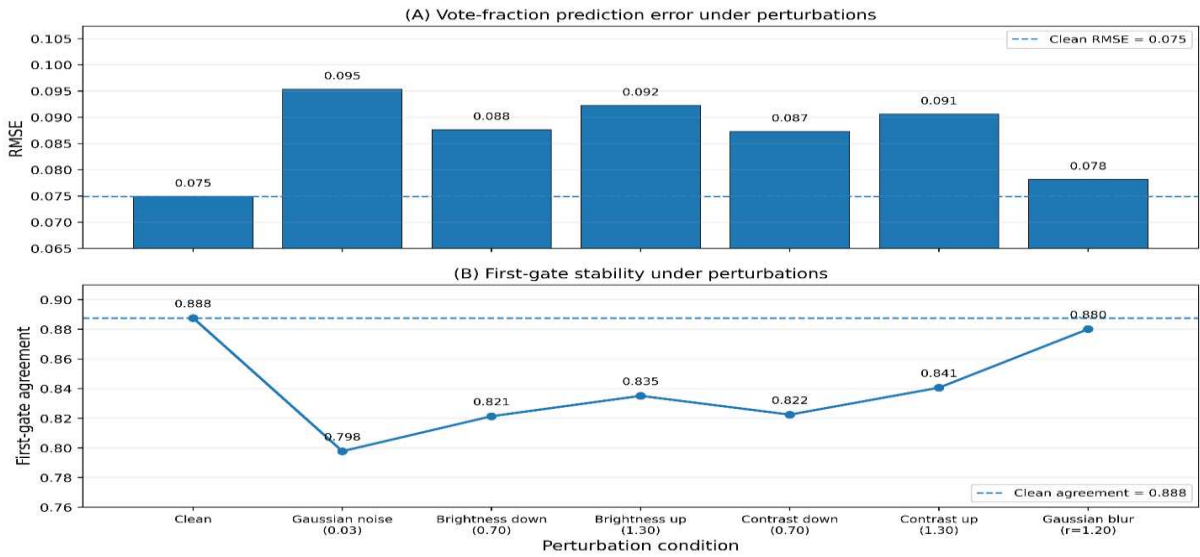


Figure 25. Framework-level robustness evidence under controlled image perturbations.

Confidence and Error Evidence for Correct and Incorrect Predictions

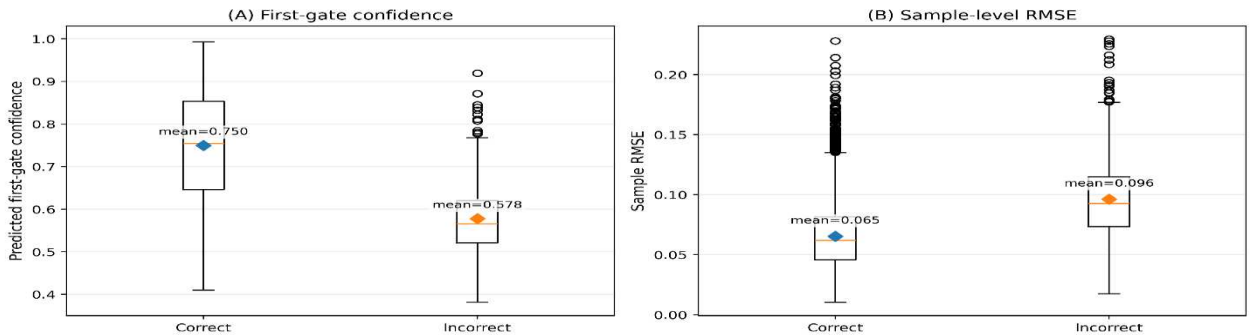


Figure 26. Confidence and error evidence for correct and incorrect galaxy morphology predictions.

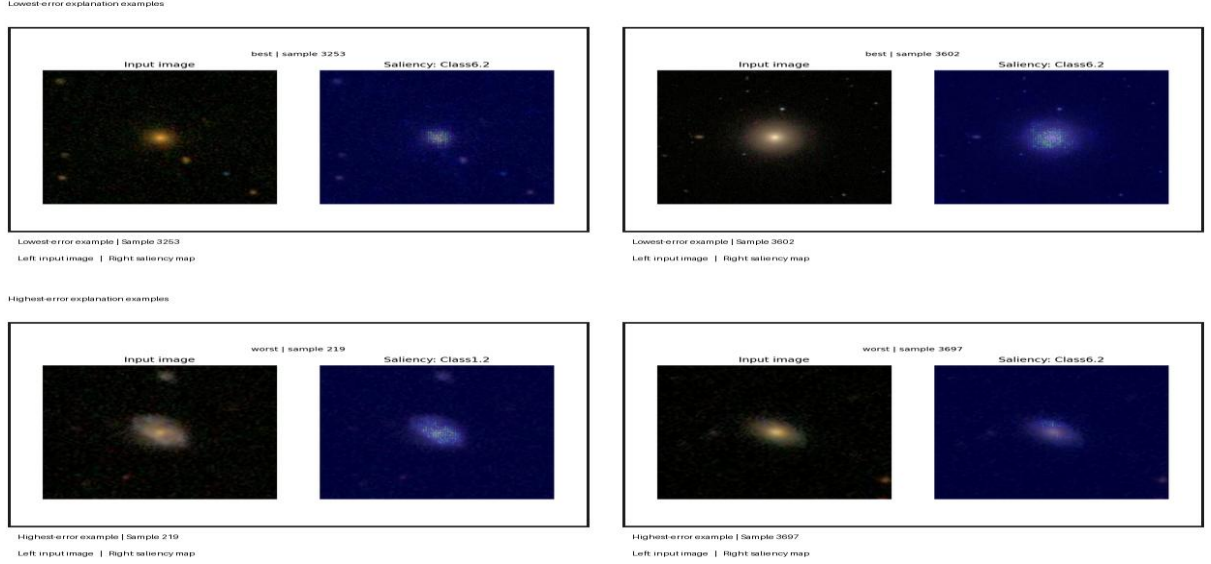


Figure 27. Saliency-based visual explanation evidence for selected lowest-error and highest-error Galaxy Zoo predictions.

4.6. Discussion

This sub-section highlights the key results from Section 4 and positions the proposed architecture in the existing literature on galaxy morphology classification. As there were two directions for evaluations, the comparative analysis is divided into two parts: Galaxy10 DECaLS hard-label classification and Galaxy Zoo 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. (2025)	Galaxy10 DECaLS classification	Semi-supervised GAN-based learning	~84% accuracy	Relevant Galaxy10 reference under limited-label learning.
Yao et al. (2024)	Galaxy10 DECaLS classification	DCGAN-based data augmentation	Up to ~95% accuracy	Strong augmentation-based Galaxy10 classification reference.
Wei et al. (2024)	Galaxy10 DECaLS classification	Vision Transformer	~94% accuracy	Strong transformer-based Galaxy10 comparison point.
Zhang et al. (2024)	Galaxy10 DECaLS classification	CNN with attention mechanism	~96% accuracy	Strong attention-based Galaxy10 classification benchmark.
This Framework	Galaxy10 DECaLS ten-class classification	ConvNeXtV2 transfer-learning CNN model	91.58% accuracy; macro F1 = 0.9073	Integrates competitive accuracy with robustness profiling and explainable AI.

Continuous vote-fraction estimation and routing performance are compared with the historical and contemporary benchmarks of surveys in Table 11.

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. (2015)	Galaxy Zoo morphology classification	Rotation-invariant CNN	~91% accuracy	Provides an early CNN-based Galaxy Zoo morphology classification benchmark.
Walmsley et al. (2022)	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. (2026)	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. (2025)	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 (Dieleman, 2014)	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%</u> <u>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: Lower values of RMSE mean better performance.

As seen from the comparative analysis, while the standalone models might be optimized for some specific metrics, the proposed framework manages to bridge hard labeling and continuous uncertainty-aware estimation. It provides the calibrated uncertainty intervals along with visual explanations, thus ensuring a solid basis for the automated morphology pipelines.

5. Conclusion

This paper proposed and assessed a confidence-aware deep learning framework for the purposes of galaxy morphology classification and analysis based on the Galaxy10 DECaLS and Galaxy Zoo dataset. Various model families were studied, including convolutional neural networks, CNN backbone of the transfer-learning models, transformers, and YOLO object detectors. The proposed framework consisted of three types of tasks: ten-class hard-label classification, high-confidence Smooth-versus-Featured routing, and prediction of 37 Galaxy Zoo vote-fractions. ConvNeXtV2 model performed with a held-out test accuracy of 91.58% and a Macro F1-score of 0.9073 on the ten-class task, while YOLO11x performed with an accuracy of 99.17% and a Macro F1-score of 0.9881 on the high-confidence binary task. As for the 37-outputs prediction, ConvNeXtV2 Large performed with an RMSE of 0.074912 and MAE of 0.042315. Additionally, the post-deadline submission on Kaggle offered extra external benchmarking evidence with the Public and Private RMSE scores of 0.07433 and 0.07440, respectively, compared to the original Galaxy Zoo Challenge winner Private RMSE score of 0.07492; however, the held-out test results were the primary ones. Along with confidence analysis, controlled robustness testing, Grad-CAM and saliency map visual explanation techniques, this work proves that the research goals have been fulfilled and the main research question was answered within the experimental scope of this study.

Nonetheless, the results of this work must be considered within the context of the used datasets and evaluation conditions. Some classes of the Galaxy10 dataset are visually similar or ambiguous, and the performed perturbations do not cover all imaging conditions or survey differences that might appear in future astronomical observations. Moreover, the visual explanation techniques provide additional evidence but not complete proof of the model reasoning. Therefore, further work on this topic might include expanding the evaluation of the framework on more astronomy surveys, developing the confidence calibration and uncertainty estimation via probabilistic and/or Bayesian approaches, performing the robustness testing under survey and observation shift

conditions, and adding the evaluation of the visual explanations quantitatively. Also, the framework can be further extended to build an AI-assisted decision support system for the initial morphology classification and explanation.

References

- Bhambra, P., Joachimi, B., and Lahav, O. (2022). “Explaining deep learning of galaxy morphology with saliency mapping”. *Monthly Notices of the Royal Astronomical Society* 511(4), pp. 5032–5041. <https://doi.org/10.1093/mnras/stac368>
- Cao, J., Wu, H., Zhang, Y., et al. (2024). “Galaxy morphology classification based on Convolutional Vision Transformer”. *Astronomy & Astrophysics* 683, A44. <https://doi.org/10.1051/0004-6361/202348544>
- Cheng, T. (2022). “Vision Transformer Applications for Galaxy Morphology Classification”. *The Astrophysical Journal* 935, Article 34. <https://doi.org/10.3847/1538-4357/ac6a52>
- Dieleman, S. (2014). “My solution for the Galaxy Zoo challenge”. *Sander Dieleman Blog*. <https://sander.ai/2014/04/05/galaxy-zoo.html>
- Dieleman, S., Willett, K. W., and Dambre, J. (2015). “Rotation-invariant convolutional neural networks for galaxy morphology prediction”. *Monthly Notices of the Royal Astronomical Society* 450(2), pp. 1441–1459. <https://doi.org/10.1093/mnras/stv632>
- Le, T., Ibrahim, N., Nguyen, T., et al. (2025). “Enhanced Comparative Analysis of Pretrained and Custom Deep Convolutional Neural Networks for Galaxy Morphology Classification”. *Engineering Proceedings* 89(1), Article 36. <https://doi.org/10.3390/engproc2025089036>
- Luo, Z., Chen, J., Chen, Z., et al. (2025). “Galaxy Morphology Classification via Deep Semi-supervised Learning with Limited Labeled Data”. *The Astrophysical Journal Supplement Series* 267(2), Article 45. <https://doi.org/10.3847/1538-4365/ad1f0c>
- 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. (2026). “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. <https://doi.org/10.1093/mnras/staf2250>
- Walmsley, M., Lintott, C., Galloway, M., et al. (2022). “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. <https://doi.org/10.1093/mnras/stab2093>
- Wei, Y., Zhang, H., and Liu, Q. (2024). “Galaxy morphology classification using Vision Transformer architectures”. *Astronomy and Computing* 48, Article 100827. <https://doi.org/10.1016/j.ascom.2024.100827>

- Xie, Y., Zhang, H., and Wang, L. (2024). “Galaxy Morphology Classification Based on VGG19 Deep Convolutional Neural Network”. *Astrophysics and Space Science*, Article 369. <https://doi.org/10.1007/s10509-024-04124-8>
- Yao, H., Liu, Y., and Zhao, D. (2024). “A Galaxy Image Augmentation Method Based on Few-shot Learning”. *Research in Astronomy and Astrophysics* 24(3), Article 035005. <https://doi.org/10.1088/1674-4527/acf92c>
- Ye, R., Shen, S., de Souza, R. S., Xu, Q., Chen, M., Chen, Z., Ishida, E. E. O., Krone-Martins, A., and Durgesh, R. (2025). “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. <https://doi.org/10.1093/mnras/staf008>
- Zhang, Y., Li, H., Chen, X., and Wang, J. (2024). “Attention-based Deep Learning Framework for Galaxy Morphology Classification”. *Astronomy and Computing* 47, Article 100810. <https://doi.org/10.1016/j.ascom.2024.100810>