

احمد الكحلوت

Deep Learning Framework for Galaxy Morphology Classification and Analysis_final.docx

 Midocean University

Document Details

Submission ID

trn:oid::2945:404713081

Submission Date

Aug 5, 2026, 6:05 AM GMT+3

Download Date

Aug 11, 2026, 12:02 PM GMT+3

File Name

Deep Learning Framework for Galaxy Morphology Classification and Analysis_final.docx

File Size

8.2 MB

83 Pages

17,387 Words

106,572 Characters

8% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Quoted Text

Match Groups

-  **127 Not Cited or Quoted 7%**
Matches with neither in-text citation nor quotation marks
-  **15 Missing Quotations 1%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 5%  Internet sources
- 5%  Publications
- 6%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Match Groups

- 127 Not Cited or Quoted 7%**
Matches with neither in-text citation nor quotation marks
- 15 Missing Quotations 1%**
Matches that are still very similar to source material
- 0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
- 0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 5% Internet sources
- 5% Publications
- 6% Submitted works (Student Papers)

Top Sources

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Internet	www.ewadirect.com	<1%
2	Submitted works	Midocean University on 2025-09-05	<1%
3	Publication	Yiqi Yao, Jinqiu Zhang, Ping Du, Shuyu Dong, Xiangru Li. "A Galaxy Image Augmen...	<1%
4	Submitted works	University of Suffolk on 2026-06-05	<1%
5	Submitted works	Manchester Metropolitan University on 2024-05-24	<1%
6	Internet	mdpi-res.com	<1%
7	Internet	sub.ifspress.hk	<1%
8	Internet	fci.stafpu.bu.edu.eg	<1%
9	Publication	Vishal Thakur, Ravindara Bhatt, R. S. Raja Durai, Satish Kumar. "Attention-Enhanc...	<1%
10	Internet	www.mdpi.com	<1%

11	Internet	hal.science	<1%
12	Internet	eprints.lib.hokudai.ac.jp	<1%
13	Internet	oulurepo.oulu.fi	<1%
14	Publication	"Web Intelligence and Human-Machine Interaction", Springer Science and Busine...	<1%
15	Publication	B. G. Nagaraja, S. Kannadhasan. "Information and Communication Systems", CRC...	<1%
16	Internet	arxiv.org	<1%
17	Internet	assets-eu.researchsquare.com	<1%
18	Internet	ijsart.com	<1%
19	Submitted works	Liverpool John Moores University on 2021-06-05	<1%
20	Publication	"Image Analysis and Processing - ICIAP 2022", Springer Science and Business Me...	<1%
21	Publication	Guangyu Gao. "Video Cataloguing - Structure Parsing and Content Extraction", CR...	<1%
22	Submitted works	Manchester Metropolitan University on 2026-05-10	<1%
23	Submitted works	VIT University on 2025-05-07	<1%
24	Internet	pure.manchester.ac.uk	<1%

25	Internet	sciencepublishinggroup.com	<1%
26	Publication	Thrinith Fernando, Samadhi Rathnayake, Kapila Dissanayaka. "Galaxy Morpholog...	<1%
27	Submitted works	University of Surrey on 2020-01-14	<1%
28	Publication	Jayme Garcia Arnal Barbedo, Marcelo Santos da Silva, Mirela Teixeira Cazzolato, L...	<1%
29	Submitted works	University of Suffolk on 2024-08-15	<1%
30	Internet	mfuir.mfu.ac.th	<1%
31	Submitted works	University of Southern California on 2012-12-06	<1%
32	Submitted works	Galgotias University, Greater Noida on 2026-05-23	<1%
33	Submitted works	Sheffield Hallam University on 2021-09-27	<1%
34	Internet	journal.perbanas.id	<1%
35	Submitted works	University of Wolverhampton on 2026-01-11	<1%
36	Internet	uvadoc.uva.es	<1%
37	Publication	"Applied Intelligence and Informatics", Springer Science and Business Media LLC,...	<1%
38	Submitted works	Cranfield University on 2026-07-07	<1%

39	Submitted works	Maulana Azad National Urdu University on 2022-01-27	<1%
40	Publication	Raul Urechiatu, Marc Frincu. "Improved Galaxy Morphology Classification with Co...	<1%
41	Publication	Chris J. Lintott, Kevin Schawinski, Anže Slosar, Kate Land et al. " Galaxy Zoo: morp...	<1%
42	Submitted works	University of Hertfordshire on 2026-01-06	<1%
43	Internet	ir.lib.nycu.edu.tw	<1%
44	Internet	portal.nersc.gov	<1%
45	Internet	qspace.qu.edu.qa	<1%
46	Publication	Goddard, Hunter. "Machine Learning Techniques for Galaxy Imagery and Photom...	<1%
47	Submitted works	R V Institute of Technology & Management, Bengaluru on 2026-05-11	<1%
48	Publication	Vaishali Vinay. "Jailbreaking Large Language Models: Techniques, Trends, Defens...	<1%
49	Internet	docslib.org	<1%
50	Internet	open.library.ubc.ca	<1%
51	Internet	repository.aaup.edu	<1%
52	Internet	repository.upnjatim.ac.id	<1%

53	Internet	www.techscience.com	<1%
54	Internet	www.universoastronomia.com	<1%
55	Publication	Augusto César de Camargo Neto. "Evaluating the robustness of AI-based vocal bi..."	<1%
56	Submitted works	Charles University in Prague on 2026-07-09	<1%
57	Publication	Handa Cai, Weibin Chen. "CLIP-GMC: Contrastive Language-image Pretraining Mo..."	<1%
58	Publication	Kam Meng Goh, Derrick Hiang Yaol Lim, Zhen Dong Sham, Kolla Bhanu Prakash. "...	<1%
59	Submitted works	Noroff University College on 2025-05-02	<1%
60	Publication	R. Gupta, P.K. Srijith, S. Desai. "Galaxy morphology classification using neural ordi..."	<1%
61	Publication	Stephen Serjeant, James Pearson, Hugh Dickinson, Johanna Jarvis. "Citizen scienc..."	<1%
62	Publication	Tanuj Sadasivam, Ria Treza Serrao, Aditya Prasad Bansod, Tanay Singh, R Bharath...	<1%
63	Submitted works	University of Hertfordshire on 2026-01-05	<1%
64	Submitted works	University of Northumbria at Newcastle on 2024-05-07	<1%
65	Internet	academic.oup.com	<1%
66	Internet	dl.futuretechsci.org	<1%

67	Internet	udspace.udel.edu	<1%
68	Internet	www.aanda.org	<1%
69	Internet	www.eecs.qmul.ac.uk	<1%
70	Internet	www.raa-journal.org	<1%
71	Publication	Antonio Di Cecco, Leonida Gianfagna. "Explainable AI with Python", Springer Scie...	<1%
72	Publication	Gabriel Marín Díaz, Alvaro Manuel Rodriguez-Rodriguez, Eva María Andrés Núñez...	<1%
73	Submitted works	Kotebe University on 2023-10-01	<1%
74	Submitted works	Midocean University on 2026-05-02	<1%
75	Publication	Ming Wei Lee, John Y.H. Soo, Syarawi M.H. Sharoni. "Classification of spiral galaxi...	<1%
76	Publication	Rahul Kumar, Md Kamruzzaman Sarker, Sheikh Rabiul Islam. "Chapter 8 Vision Tr...	<1%
77	Publication	Sai Teja Erukude, Lior Shamir. "Galaxy image simplification using Generative AI", ...	<1%
78	Submitted works	Technical University of Košice on 2026-06-03	<1%
79	Submitted works	The University of Manchester on 2022-11-06	<1%
80	Submitted works	UCL on 2025-02-24	<1%

81	Submitted works	University of Gujrat on 2026-06-17	<1%
82	Submitted works	Vilnius Gediminas Technical University on 2026-05-29	<1%
83	Internet	ir.unisa.ac.za	<1%
84	Internet	nzjohng.github.io	<1%
85	Submitted works	Robert Kennedy College on 2026-07-07	<1%
86	Submitted works	Universiti Tun Hussein Onn Malaysia on 2026-05-25	<1%
87	Submitted works	University of Auckland on 2011-10-18	<1%
88	Submitted works	University of Hertfordshire on 2025-01-06	<1%
89	Submitted works	University of New South Wales on 2026-05-01	<1%
90	Publication	V. Sekhar, P. Ajay Kumar Reddy, K. Logesh, T. Raghunadha Reddy, Dara Raju. "Ne...	<1%
91	Submitted works	Walsh College on 2026-03-24	<1%
92	Submitted works	Leiden University on 2022-06-30	<1%
93	Submitted works	Manipal Academy of Higher Education (MAHE) on 2024-08-07	<1%
94	Publication	Sanchari Dhar, Lior Shamir. "Systematic biases when using deep neural networks...	<1%

95	Publication	Slijepcevic, Inigo Val. "Advancing Representation Learning for Computer Vision in..."	<1%
96	Submitted works	Sunway Education Group on 2025-05-25	<1%
97	Submitted works	University of Surrey on 2024-01-11	<1%
98	Internet	ir.nctu.edu.tw	<1%
99	Submitted works	Asian Institute of Technology on 2026-05-04	<1%
100	Internet	prod-ms-be.lib.mcmaster.ca	<1%



Deep Learning Framework for Galaxy Morphology Classification and Analysis

Prepared By

Name: Ahmad Yousef Alkahlout

ID No.: 253000753

Name: Mohammed Taleb Ragash Mohammed

ID No.: 252000097

Submitted in fulfillment of the requirements for the degree Master of
Science in Artificial Intelligence

At

College of Informatics

Midocean University

Supervisor Name: Dr. Mohamed Essa

College of Informatics, Midocean University, Comoros

Email: 253000753@st.midocean.edu.km /
252000097@st.midocean.edu.km

Acknowledgements

50 We would like to thank our supervisor, Dr. Mohamed Essa, for his continuous guidance and valuable feedback. His comments challenged us to put the maximum efforts and to ensure success on this research.

81 We are also thankful to the faculty members at the College of Informatics, Midocean University, for providing a supportive academic environment that inspired independent thinking and research development.

25 Finally, we would like to express our appreciation to our families for their patience, encouragement, and support during this journey.

Abstract

Classification of galaxy morphology is one of the fundamental problems in astrophysics as galaxy appearance carries vital information about its origin and evolution. The increasing number of imaged galaxies due to various survey missions, such as SDSS and DECaLS, makes a human classification impossible for any large-scale research. Even though the deep learning algorithms perform exceptionally well in terms of classification quality, a lot of studies evaluate only the predictive accuracy of the model neglecting the issues of robustness, uncertainty, and interpretability.

In our thesis, we propose a confidence-aware deep learning framework for galaxy morphology classification and analysis. The model types that were evaluated in this study include baseline CNNs, group-equivariant CNNs, transfer-learning CNNs, transformer-based and hybrid models, YOLO classifiers, and ensembling approaches. Galaxy10 DECaLS dataset was used for ten-class hard label galaxy morphology classification and Galaxy Zoo Challenge dataset was used for 37 continuous output vote fraction predictions. Moreover, we developed a high-confidence binary Smooth-versus-Featured classification approach based on Galaxy Zoo first-gate vote fractions. For classification evaluation, accuracy, precision, recall, and Macro F1-score were used, while RMSE and MAE were used for continuous vote-fraction prediction. Model robustness was also evaluated with controlled image perturbations, confidence analysis, Grad-CAM and saliency maps.

For the Galaxy10 DECaLS test dataset, the best result was obtained by the ConvNeXtV2 transfer-learning model, with 91.58% accuracy and a Macro F1-score of 0.9073. The best results for high-confidence Smooth-versus-Featured classification task were reached by YOLO11x classifier with 99.17% accuracy and 0.9881 Macro F1-score. For Galaxy Zoo 37-output prediction task, ConvNeXtV2 Large model achieved RMSE of 0.074912 and MAE of 0.042315.

Table of Contents

Acknowledgements	ii
Abstract	iii
List of Abbreviations	vi
List of Figures	vii
List of Tables	ix
Chapter 1: Introduction	1
1.1 OVERVIEW	1
1.2 PROBLEM DEFINITION	3
1.3 THESIS OBJECTIVES	4
1.4 THESIS CONTRIBUTION	5
1.5 THESIS ORGANIZATION	6
Chapter 2: Background and Related Work	7
2.1 INTRODUCTION	7
2.2 BACKGROUND	7
2.3 RELATED WORK	12
2.4 RELATED WORK SUMMARY	19
Chapter 3: Proposed Methodology	23
3.1 INPUT (DATASETS)	24
3.1.1 <i>Galaxy10 DECaLS Dataset</i>	24
3.1.2 <i>Galaxy Zoo Challenge Dataset</i>	25
3.2 PREPROCESSING	27
3.2.1 <i>Preprocessing for Hard-Label Classification</i>	27
3.2.2 <i>Preprocessing for Vote-Fraction Hierarchical Classification</i>	28
3.3 EXPERIMENTAL CONTROL AND DATA PARTITIONING	28
3.4 CLASSIFICATION MODELS	29
3.4.1 <i>Deep Learning CNN Family</i>	31
3.4.2 <i>Transfer-Learning CNN Backbones Family</i>	33
3.4.3 <i>Transformer and Hybrid Attention Family</i>	35
3.4.4 <i>YOLO Classification Family</i>	37
3.4.5 <i>Ensemble and Hierarchical Classification Strategy</i>	39
3.5 TRAINING AND EVALUATION	41
Chapter 4: Experimental Results and Discussion	43
4.1 PREPROCESSING RESULTS	43

4.2 HARD-LABEL CLASSIFICATION RESULTS ACROSS MODEL FAMILIES.....	45
4.2.1 Detailed Analysis of the ConvNeXtV2 Transfer-Learning CNN Model...	46
4.3 HIGH-CONFIDENCE SMOOTH VS FEATURED CLASSIFICATION.....	51
4.4 GALAXY ZOO VOTE-FRACTION PREDICTION RESULTS.....	57
4.5 RELIABILITY-ORIENTED EVALUATION.....	62
4.6 DISCUSSION.....	65
Chapter 5: Conclusion and Future Work.....	70
References.....	72

List of Abbreviations

- AI: Artificial Intelligence.
- AUC: Area Under the Curve.
- BASS: Beijing–Arizona Sky Survey.
- C2f: Cross Stage Partial Bottleneck with Two Convolutions.
- CANDELS: Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey.
- CNN: Convolutional Neural Network.
- CSP: Cross Stage Partial.
- CvT: Convolutional Vision Transformer.
- DECaLS: Dark Energy Camera Legacy Survey.
- DESI: Dark Energy Spectroscopic Instrument.
- DL: Deep Learning.
- FSL-GAN: Few-Shot Learning–Generative Adversarial Network.
- GAN: Generative Adversarial Network.
- GCNN: Group-Equivariant Convolutional Neural Network.
- Grad-CAM: Gradient-weighted Class Activation Mapping.
- GZD-5: Galaxy Zoo DECaLS 5.
- MAE: Mean Absolute Error.
- MBConv: Mobile Inverted Bottleneck Convolution.
- MzLS: Mayall z-band Legacy Survey.
- RMSE: Root Mean Squared Error.
- S0: Lenticular Galaxy.
- S-PLUS: Southern Photometric Local Universe Survey.
- SDSS: Sloan Digital Sky Survey.
- SE: Squeeze-and-Excitation.
- ViT: Vision Transformer.
- VGG19: Visual Geometry Group 19-layer Network.
- XAI: Explainable Artificial Intelligence.
- YOLO: You Only Look Once.

List of Figures

Figure 1.1 Example of galaxy morphology complexity relevant to automated galaxy classification studies.....	1
Figure 2.1 Hubble morphological classification scheme illustrating elliptical, spiral, and barred spiral galaxies.	8
Figure 2.2 Example galaxy images from different morphological classes used in automated classification studies.....	10
Figure 2.3 Schematic overview of a deep learning pipeline for galaxy morphology classification.....	11
Figure 2.4 Example of a saliency map highlighting discriminative regions in galaxy morphology classification.....	14
Figure 3.1 Overall proposed galaxy morphology methodology framework.	23
Figure 3.2 Sample galaxy images from the Galaxy10 DECaLS dataset used for hard-label morphology classification.....	25
Figure 3.3 Sample Galaxy Zoo Challenge images with first-gate vote-fraction annotations used for hierarchical morphology classification.....	26
Figure 3.4 CNN family architecture used for local galaxy morphology feature extraction.	32
Figure 3.5 Transfer-learning CNN backbone family adapted for galaxy morphology classification.....	35
Figure 3.6 Transformer and hybrid attention architecture family used for broader spatial morphology reasoning.	37
Figure 3.7 YOLO classification architecture family used as an efficient image-level morphology classifier	38
Figure 3.8 Ensemble and hierarchical prediction strategy connecting model families with 37-output Galaxy Zoo morphology prediction.....	41
Figure 4.1 Representative source and preprocessed galaxy images from the Galaxy10 DECaLS and Galaxy Zoo datasets.....	44
Figure 4.2 Galaxy10 DECaLS ten-class confusion matrix on the held-out test set.	48
Figure 4.3 Representative correct and incorrect Galaxy10 DECaLS predictions from the ConvNeXtV2 transfer-learning CNN model.	48
Figure 4.4 Grad-CAM visual explanations for selected Galaxy10 DECaLS predictions from the ConvNeXtV2 transfer-learning CNN model.	49

Figure 4.5 Robustness evaluation under controlled image perturbations.	50
Figure 4.6 Uncertainty distribution for correct and incorrect Galaxy10 DECaLS predictions.	51
Figure 4.7 Galaxy Zoo first-gate vote-fraction examples used to construct the high-confidence Smooth vs Featured binary classification task.	52
Figure 4.8 Official held-out test confusion matrix for the high-confidence Smooth vs Featured binary classification stage.....	54
Figure 4.9 Representative correct and incorrect predictions from the high-confidence Smooth vs Featured classification stage.	55
Figure 4.10 Full-test and selective-confidence performance of the Smooth vs Featured classification stage.	56
Figure 4.11 Robustness of the Smooth vs Featured classifier under controlled image perturbations.	57
Figure 4.12 Per-output RMSE values for the Galaxy Zoo 37-output vote-fraction prediction task.	59
Figure 4.13 First-gate prediction analysis for the Galaxy Zoo 37-output vote-fraction model.....	60
Figure 4.14 Representative lowest-error and highest-error Galaxy Zoo vote-fraction prediction examples.....	61
Figure 4.15 Framework-level robustness evidence under controlled image perturbations.	63
Figure 4.16 Confidence and error evidence for correct and incorrect galaxy morphology predictions.	64
Figure 4.17 Saliency-based visual explanation evidence for selected lowest-error and highest-error Galaxy Zoo predictions.	65

List of Tables

Table 2.1 Comparison of related studies in galaxy morphology classification.	19
Table 3.1 Galaxy10 DECaLS split used in the main classification experiments.	24
Table 3.2 Galaxy Zoo Challenge labeled split and official unlabeled test set role.	26
Table 4.1 Preprocessing outputs used in the framework.	44
Table 4.2 Hard-label classification performance across the evaluated model families.	45
Table 4.3 Per-class classification report for the ConvNeXtV2 transfer-learning CNN model on the Galaxy10 DECaLS ten-class test set.	47
Table 4.4 Per-class classification report for the YOLO11x high-confidence Smooth vs Featured classifier on the held-out test set.	53
Table 4.5 Official Galaxy Zoo 37-output vote-fraction prediction results on the held-out test set.	58
Table 4.6 Reliability evidence supporting the proposed galaxy morphology framework.	62
Table 4.7 Comparison with Galaxy10 DECaLS hard-label classification studies.	66
Table 4.8 Comparison with Galaxy Zoo and Kaggle benchmark evidence.	67

Chapter 1: Introduction

1.1 Overview

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].

As shown in Figure 1.1, 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.



Figure 1.1 Example of galaxy morphology complexity relevant to automated galaxy classification studies.
Source: Adapted from NASA Astronomy Picture of the Day [21]. Image credit: NASA, ESA, Hubble, and HLA; processing by Domingo Pestana.

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 was 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 thesis to a certain extent. It should be noted that the current thesis 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 thesis is addressed from the angle of performance, reliability, and interpretability [13].

1.2 Problem Definition

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 a 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 others may contain features corresponding to several classes. As a result, labeling only by final class may not be 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 thesis 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.

1.3 Thesis Objectives

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

- Create baseline CNN models to classify galaxy morphology and assess their capability of learning the features directly from the images.
- Use a 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 a 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 assess 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.

1.4 Thesis 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.

1.5 Thesis Organization

Further organization of this thesis is as follows. Chapter 2 will provide an overview of the theory and literature on related research fields. Chapter 3 gives a description of the methodology. Chapter 4 discusses the results obtained from the experiment. Finally, chapter 5 presents the conclusions.

Chapter 2: Background and Related Work

2.1 Introduction

This chapter gives a brief introduction of the background information and previous works about galaxy morphology classification using deep learning approaches. First, this chapter introduces some basic terminology 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 Chapter 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 the accuracy of galaxy morphology 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 chapter 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.

2.2 Background

Galaxy morphology studies date back to the early 20th century when Hubble proposed one of the first galaxy classification schemes [4]. According to this classification, galaxies were divided into several basic morphological types like

elliptical, spiral, and irregular galaxies. Hubble's contribution laid the groundwork for further exploration and research of galaxy morphology and development.

An illustrative representation of the Hubble galaxy classification scheme can be found in Figure 2.1. It illustrates the connection between elliptical galaxies and the two basic branches of spiral galaxies – normal spirals and barred spirals. Later, de Vaucouleurs elaborated Hubble's classification by introducing more categories, including S0 lenticular galaxies and other morphological distinctions [5].

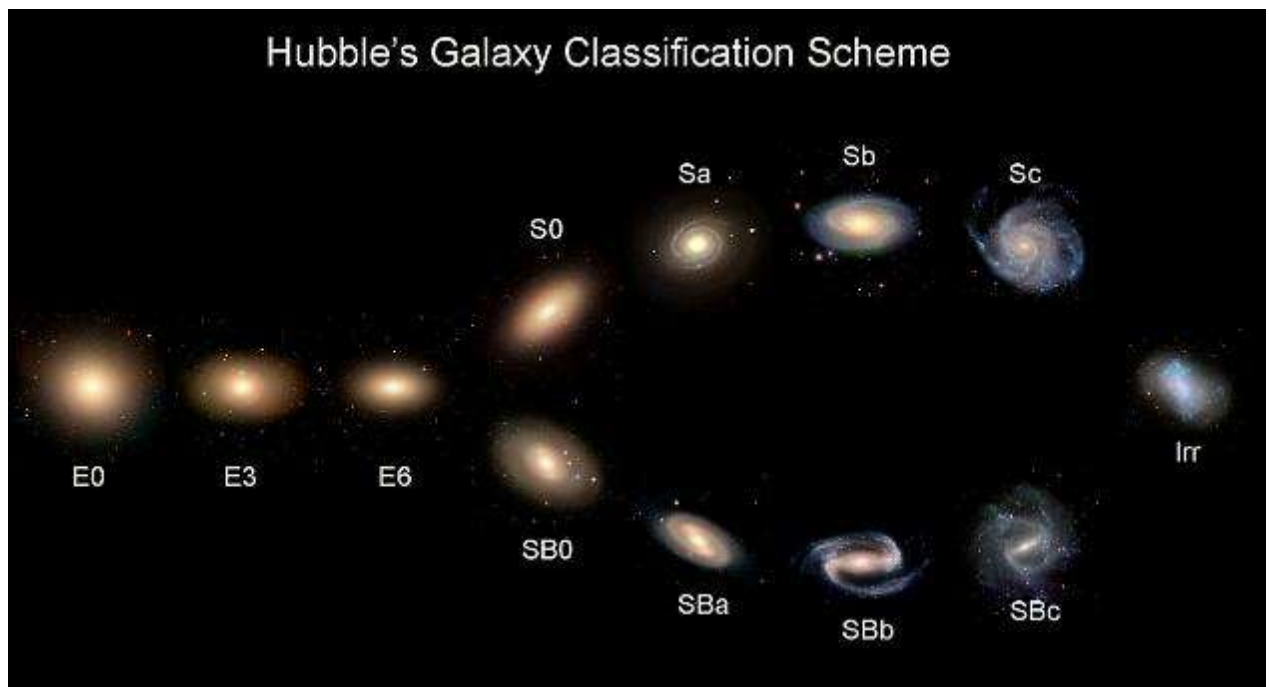


Figure 2.1 Hubble morphological classification scheme illustrating elliptical, spiral, and barred spiral galaxies.
Source: Adapted from [Hubble \(1936\) \[4\]](#).

For many years, the process of galaxy classification was based on the manual visual analysis of galaxy images until relatively recently. It relied on the judgement of astronomers who identified the basic structural features of galaxies and assigned them particular classification labels. Even though this traditional method of galaxy classification remains valid to some extent, it became rather impractical due to the increasing number of galaxy observations.

92
41
The issue of manually classifying large amounts of galaxy images became evident with the rise of modern surveys. The Sloan Digital Sky Survey (SDSS) produced a large quantity of images that could not be classified manually easily and efficiently [6]. As an alternative solution, the Galaxy Zoo project was initiated. It provided an opportunity for the public to participate in galaxy classification. It demonstrated that citizen science could be used to classify large volumes of galaxy images.

One of the most valuable aspects of a dataset such as Galaxy Zoo is that it is able to provide not just information about how something has been visually classified, but also information on how much consensus there was amongst the classifiers through use of vote fractions. This is beneficial for studying galaxy morphology, because some galaxies do not fit into one particular category. Instead, their morphology may show signs of being a blend or mixture.

93
26
As mentioned earlier, there has been a steady increase in the volume of astronomical data due to survey projects like the Dark Energy Camera Legacy Survey (DECaLS). Thus, the problem of automatic galaxy classification has become urgent, and researchers started employing deep learning approaches to solve it.

20
Currently, several approaches including artificial intelligence, machine learning, and deep learning have been actively used in the sphere of astronomical image analysis. Machine learning helps to train models via available training data, while deep learning includes the application of multilayered neural networks for pattern extraction. These techniques can be beneficial in solving the problem of galaxy morphology classification due to the presence of numerous visual characteristics that cannot be coded manually. Indeed, as can be seen from Figure 2.2, there is a large visual diversity of galaxies that makes automated classification rather challenging.

87
Figure 2.2 gives examples of galaxy images from diverse morphological classes. These examples show the high degree of variability of visual characteristics, thus making the task of classification a difficult one for a computer vision algorithm.

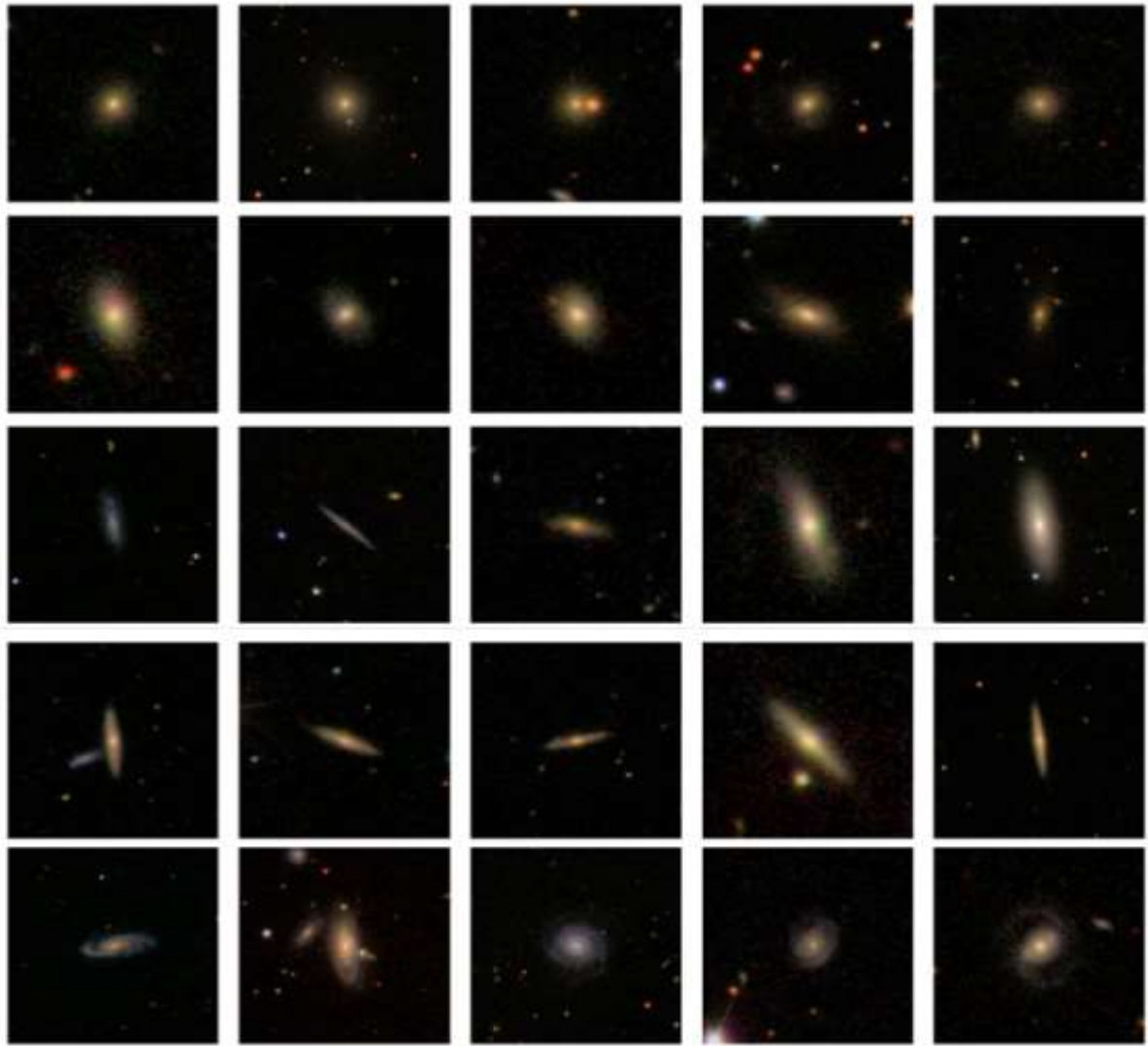


Figure 2.2 Example galaxy images from different morphological classes used in automated classification studies.
Source: Adapted from [Zhu et al. \[19\]](#).

As follows from the above discussion, the problem of automated galaxy classification can be viewed as a typical problem of image classification. Therefore, in order to detect specific galaxy morphological features, the model should learn to see particular image patterns. In this regard, the application of deep learning approaches, and particularly CNNs, may prove effective since such models allow detecting specific visual structures in images.

In modern approaches to galaxy morphology classification, many researchers use CNN-based models due to their superior performance. For instance, Dieleman et al. found that such models can provide classification accuracy at the level of

human performance. Thus, CNNs showed quite promising results when employed for predicting galaxy morphologies in [2].

Figure 2.3 shows the basic steps involved in creating a model for galaxy morphology classification using a deep learning approach. This figure demonstrates the process of classifying the galaxy's image by applying a sequence of learning operations from input to output.

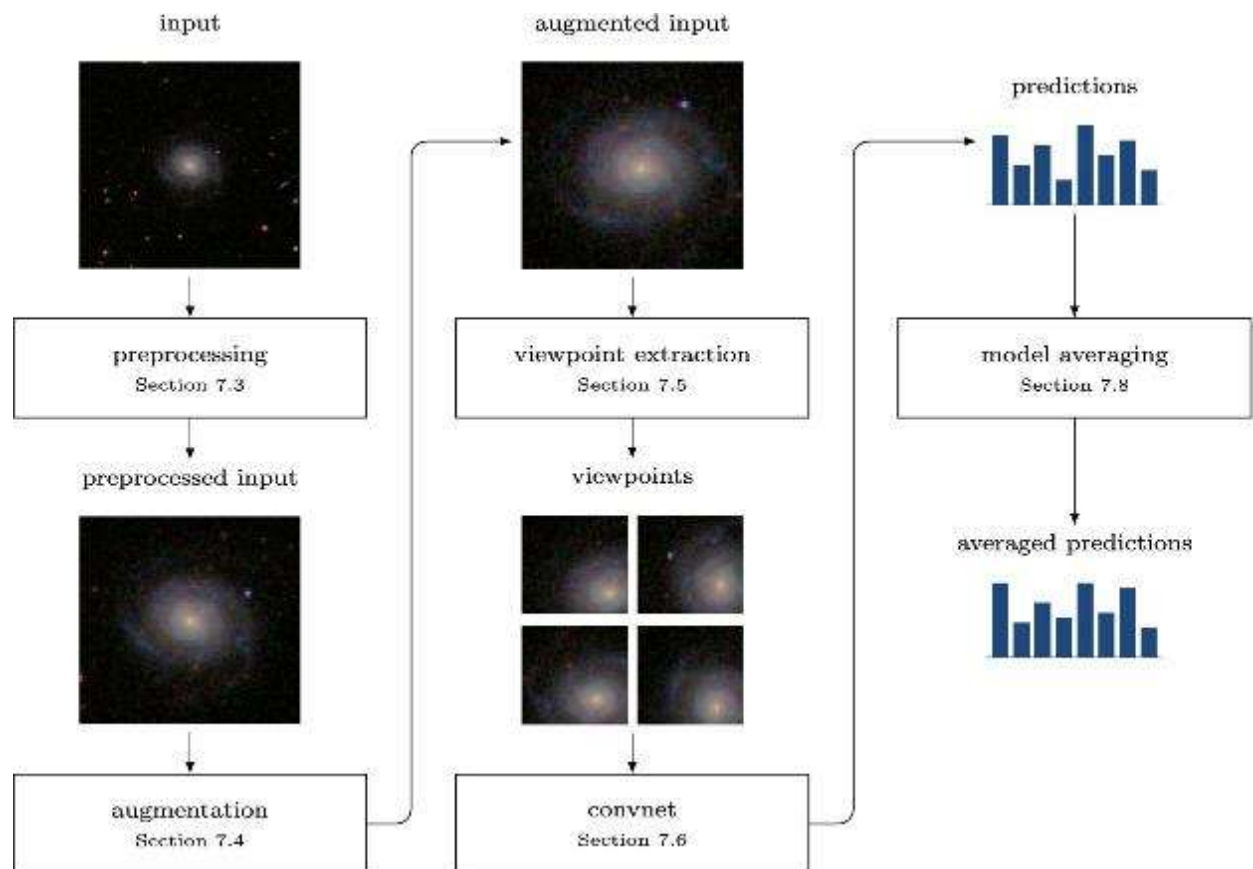


Figure 2.3 Schematic overview of a deep learning pipeline for galaxy morphology classification.

Source: Adapted from Dieleman et al. [2].

Apart from the use of CNNs in the process of galaxy morphology classification, researchers have developed other approaches to improve the accuracy of models. One such approach is related to the use of transfer learning methods for fine tuning pre-trained models on galaxy images and improving their performance in cases where labeled data are few. Transformer-based models are another direction that allows detecting more complicated visual structures. Some researchers also

apply the semi-supervised approach to learning in order to rely less on the labeled data. All this means that currently used methods of galaxy morphology classification do not limit themselves to one solution but seek to develop new models based on varied approaches.

At present, the development of computer vision algorithms goes beyond simple CNN approaches to more sophisticated architectures of visual recognition systems. The fact is that in galaxy morphology classification, there are visual features that can be detected locally while other ones require analyzing the whole image structure. Thus, modern vision architectures can be considered as a good direction in building classification models that meet the requirements of complex astronomical imagery.

Recently, researchers paid considerable attention to model interpretability issues. XAI has become widely adopted in many scientific areas. The main advantage of explainable AI techniques is the possibility of identifying image parts affecting the model's decision-making processes [3].

2.3 Related work

The reviewed literature indicates that there have been great strides made in the field of galaxy morphological classification. Nonetheless, it is apparent that the 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. \[1\]](#) 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 thesis 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. \[2\]](#) propose a deep learning algorithm for morphology prediction using a rotation-invariant CNN with model averaging. With 61,578 Galaxy Zoo 2 images as input, the goal of this algorithm is to predict 37 vote fraction features on galaxy morphologies. This approach works by rotating, scaling, and mirroring every galaxy image and averaging the predictions from various transformed images and architectures to avoid dependence on image orientation. The algorithm gives an RMSE of 0.07492 on the private leaderboard and more than 99% accuracy for most of the morphology questions if there is high agreement among the volunteers. One strength of this algorithm is that it can take advantage of rotation invariance and generate morphology answers directly from image data. The drawback is that averaging predictions increases computational cost.

[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.

As shown in Figure 2.4, 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 thesis, a similar explainability approach was used to evaluate whether the proposed deep learning framework focuses on relevant morphological structures when classifying galaxy images.

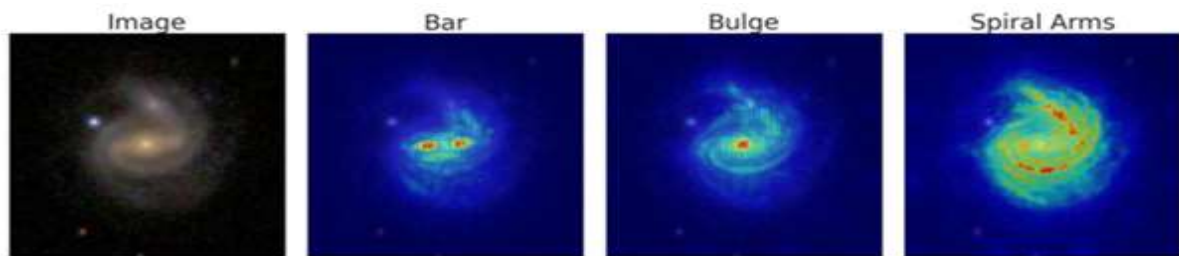


Figure 2.4 Example of a saliency map highlighting discriminative regions in galaxy morphology classification.

Source: Adapted from [Bhambra et al. \[3\]](#).

[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 Domínguez Sánchez 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 where a limited number of labeled samples are available, which is common in 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.

3 [Yao et al. \[9\]](#) present a galaxy image augmentation system that combines few-shot learning techniques with generative adversarial networks to address the problem of insufficient numbers of labeled images of galaxies. They use the Galaxy10 DECaLS dataset and test the FSL-GAN system on three classification models – AlexNet, VGG, and ResNet. According to the findings, the biggest increase in quality is achieved when the number of images in the generated dataset varies between 0.5 and 1.0 of the initial dataset. Applying the FSL-GAN technique to the ResNet model produces F1-scores equal to 89.54% for the class of In-between Round Smooth Galaxies and 63.18% for Unbarred Loose Spiral Galaxies. One of the strong points of the paper is the consideration of the effect of different augmentation techniques on various galaxy classes and classification models. Still, there is no one-size-fits-all augmentation technique that works well for all morphological classes of galaxies.

74 [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.

46 [Fernando et al. \[11\]](#) conducted research on galaxy morphology classification with a deep convolutional neural network based on the VGG19 architecture. As the data source, the authors used the Galaxy10 DECaLS dataset that contains

17,736 images that are divided into ten morphological classes. Transfer learning is conducted via using VGG19 architecture as the feature extraction backbone, along with fully connected layers, image preprocessing, and data augmentation. The performance of the suggested approach was measured via accuracy (97.97%), Macro Precision (89.84%), Macro Recall (89.69%), Macro F1-score (89.68%), and AUC (0.98). One of the main advantages of the conducted research is that it proves the effectiveness of transfer learning in the case of ten-class galaxy morphology classification. However, there are some disadvantages of the research related to the class imbalance in Galaxy10 DECaLS and computational complexity of VGG19.

48 [Parker et al. \[12\]](#) present AstroCLIP, a cross-modal foundation model developed for analysis of galaxy images and optical spectroscopy. The researchers used 77 imaging data obtained by the DESI Legacy Imaging Survey and spectral data collected by the Dark Energy Spectroscopic Instrument, using Galaxy Zoo DECaLS classifications for evaluation of morphology classification. The model is based on self-supervised transformer encoders and employs contrastive learning for alignment of image and spectral representations in a common latent space. Within the problem of morphology classification, AstroCLIP managed to demonstrate an average accuracy of around 77% in answering ten questions from Galaxy Zoo regarding morphology with the best result being 97% in disk-edge-on case. One of the main advantages of the model is its transferability to various astronomical tasks requiring only minor supervision. On the other hand, the approach requires large pretraining and significant computational power, and its results are rather poor in the case of complex questions requiring multiple classes.

40 [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 40 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.

18 Urechiatu et al. [14] focused on the potential usefulness of convolutional neural
51 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 96.83%. However, despite high effectiveness in recognizing simple structures of galaxies, the models failed to work properly when dealing with complicated morphologies.

19 Bom et al. [15] provide a deep learning-based catalog of galaxy morphology
19 using images from the S-PLUS Data Release 3. The researchers use a convolutional neural network for classifying galaxies into either early or late types and develop an additional classifier that can detect images that are of low quality. The suggested method has reached 98.5% purity rate in differentiating between spiral and elliptical galaxies. The developed catalog includes 164,314 images, 55,000 of which are of reliable galaxies. One of the strong points of the research is the combination of deep learning and large-scale survey in astronomy as well as automatic evaluation of the quality of the images. Yet, the classification problem is mostly solved through early and late galaxy types and not detailed multi-classification.

70 Mao et al. [16] The DenseNet-SE4 model was developed for improving the
24 discrimination of morphologies of visually similar galaxies. The authors created a dataset containing eight classes of galaxies taken from the Galaxy10 DECaLS and GZD-5 datasets. This new architecture is an expansion of the DenseNet-121 model based on the use of dynamic multiscale convolution, learnable grouped convolution, and squeeze-and-excitation attention layers to extract more features. This architecture showed 88.96% accuracy, 89.00% precision, 89.44% recall, and 89.21% F1-score. The main advantage of this research is that the model can

distinguish small differences between morphologically similar classes of galaxies and show good results at different redshifts. However, the number of classes used in the dataset, eight, differs from the original number of classes, ten, in the Galaxy10 DECaLS dataset.

Ye et al. [17] explored the effectiveness of unsupervised domain adaptation techniques for improving galaxy morphology classification between various surveys. The experiment utilized the labeled images from the Galaxy Zoo DECaLS database as the source domain, while the unlabeled images were taken from the BASS and MzLS surveys and considered the target domain. At first, the algorithm was trained using the DECaLS images and then it was fine-tuned using the BASS/MzLS images to eliminate the gaps that were caused by different resolutions, noise levels, and surveys. The fine-tuned algorithm showed accuracy results between 79.5% and 96.7% for the morphology questions and 87.5% accuracy for the Smooth or Featured question. One of the major benefits of the approach described above is that it can be used to transfer galaxy morphology classification labels to another survey without any labeled target-domain images. However, this approach relies on the assumption that both galaxy samples have identical underlying distributions.

Chen et al. [18] propose a semi-supervised learning algorithm for classification of galaxies using diversity and uncertainty-based learning strategies. They test their proposed framework on the Galaxy10 DECaLS and Galaxy Zoo 2 benchmarks using a few labeled examples in combination with many unlabeled examples. They vary their training procedure depending on the level of confidence in the pseudo-labels and include uncertainty in their approach to avoid confirmation bias due to high-confidence predictions. On the Galaxy10 DECaLS benchmark, their algorithm achieves 76.8% accuracy with only 2,000 labeled examples, whereas the supervised baseline yields 70.8% accuracy on the same few labeled data condition. An important contribution of this study is the improved utilization of unlabeled examples and consideration of uncertainty in predictions. However, the results presented in this paper were obtained through a limited-label setting experiment and cannot be compared directly with the fully supervised evaluation results obtained on the whole labeled Galaxy10 DECaLS benchmark.

2.4 Related Work Summary

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 2.1.

Table 2.1 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	Private RMSE = 0.07492; >99% accuracy for most high-agreement questions	Galaxy Zoo	Direct benchmark for 37-output vote-fraction 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	FSL-GAN Data Augmentation	F1-score up to 89.54%	Galaxy10 DECaLS	Data augmentation reference for limited and imbalanced galaxy classes
[10]	Le et al.	2025	EfficientNetB5 Transfer Learning	~87%	Galaxy10 DECaLS	Transfer learning architecture benchmark
[11]	Fernando et al.	2024	VGG19 Transfer Learning	~97.97%	Galaxy10 DECaLS	Transfer-learning benchmark for ten-class galaxy morphology classification
[12]	Parker et al.	2024	Self-supervised Vision Transformer / AstroCLIP	~44% – ~97% across morphology questions	Galaxy Zoo DECaLS	Recent transformer-based reference for self-supervised and confidence-aware morphology analysis
[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)	~96.83%	Galaxy Zoo 2	CNN-based classification improvement reference
[15]	Bom et al.	2024	CNN with image-quality assessment	98.5% purity	S-PLUS Data Release 3	Large-survey binary morphology catalogue and automated image-quality reference

[16]	Mao et al.	2025	DenseNet-SE4 with multiscale convolution and attention	~88.96%	Galaxy10 DECaLS and GZD-5	Attention-enhanced CNN reference for visually similar morphology classification and robustness analysis
[17]	Ye et al.	2025	Unsupervised Domain Adaptation	79.5%–96.7% across morphology questions	Galaxy Zoo DECaLS and BASS/MzLS	Cross-survey domain-adaptation reference for robustness under survey shift
[18]	Chen et al.	2025	Semi-supervised learning with uncertainty estimation	~76.8% with 2,000 labeled samples	Galaxy10 DECaLS	Uncertainty-aware semi-supervised reference under limited labels

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 binary classification approach in data on galaxy morphology 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.

Chapter 3: Proposed Methodology

This chapter describes the methodological approach used to create the proposed galaxy morphology classification and analysis system. Specifically, this chapter 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 chapter.

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 3.1 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.

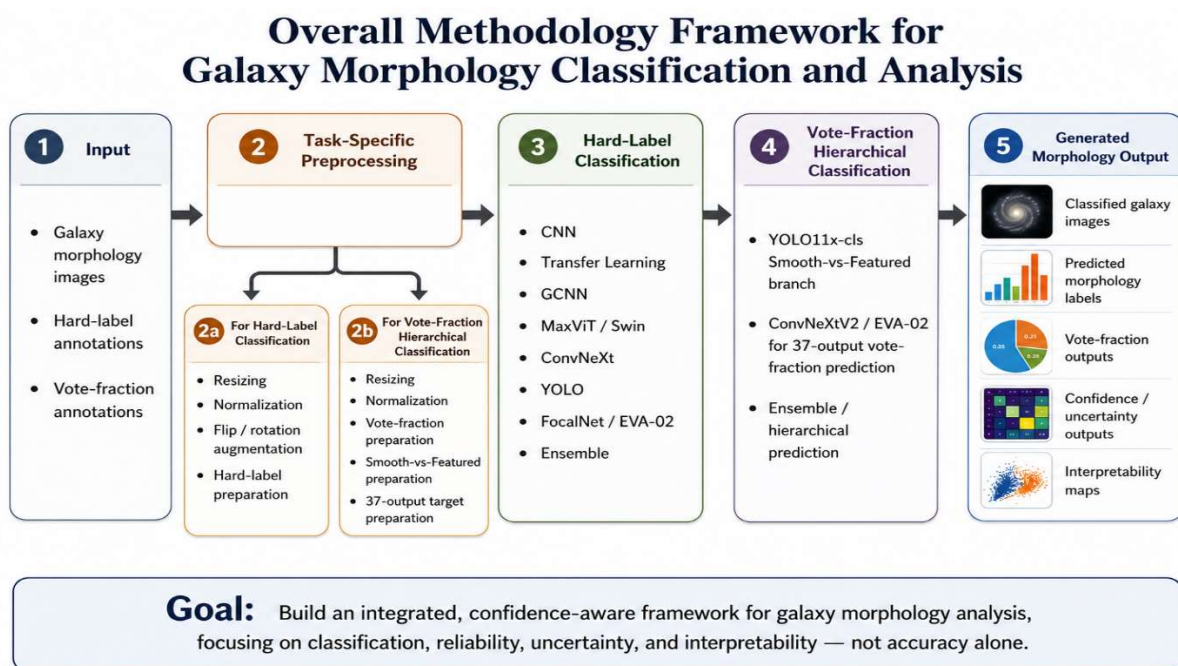


Figure 3.1 Overall proposed galaxy morphology methodology framework.

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 3.1 summarizes the training, validation, and held-out test subsets used in the main Galaxy10 DECaLS classification experiments.

Table 3.1 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 3.2 provides examples of images from the Galaxy10 DECaLS dataset illustrating the variety of hard-label galaxy morphology classification task analyzed in this study.

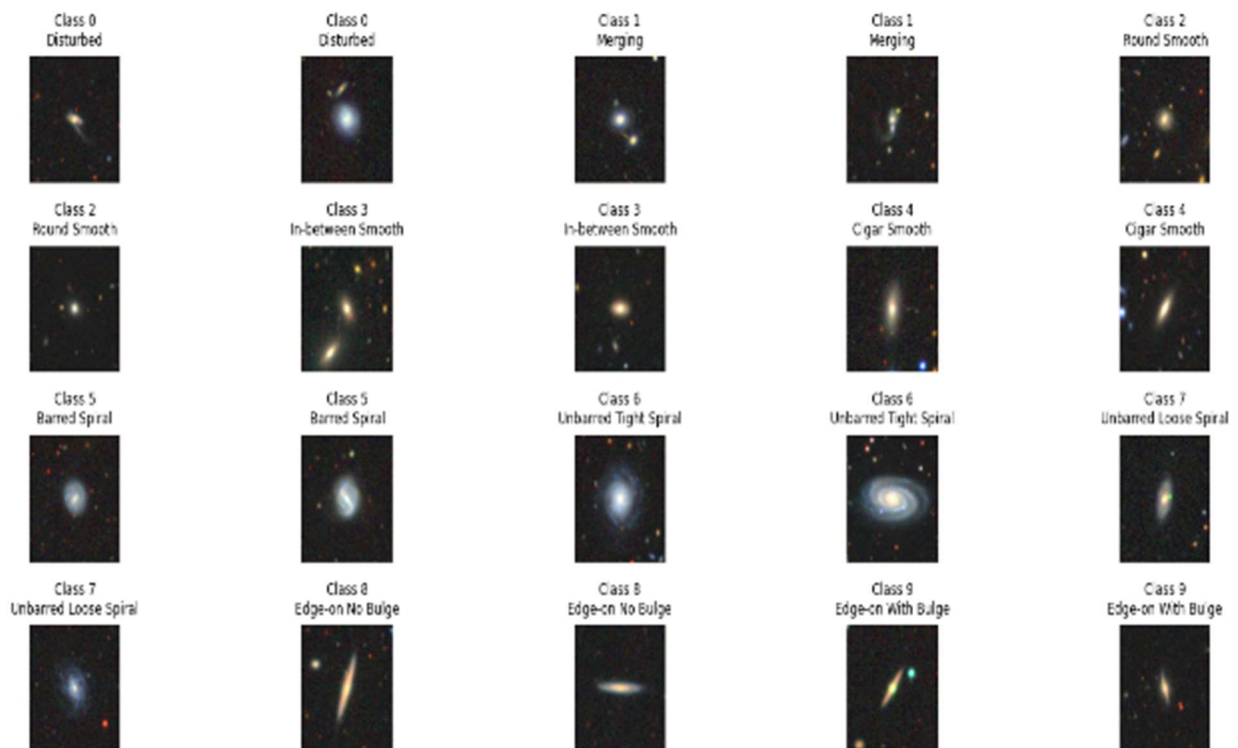


Figure 3.2 Sample galaxy images from the Galaxy10 DECaLS dataset used for hard-label morphology classification.

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 3.2 summarizes the training, validation, and internal test subsets of the labeled Galaxy Zoo Challenge dataset and clarifies the role of the official unlabeled Kaggle test set.

Table 3.2 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 3.3 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.

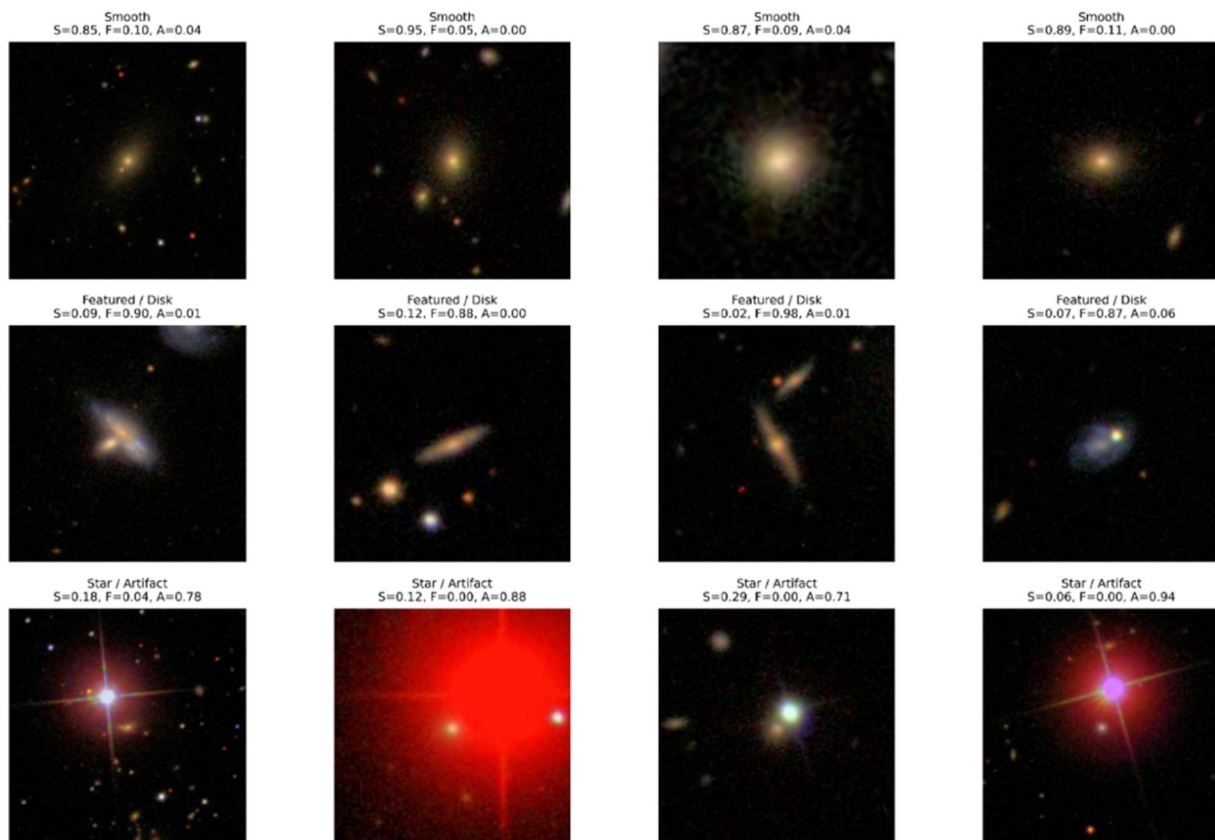


Figure 3.3 Sample Galaxy Zoo Challenge images with first-gate vote-fraction annotations used for hierarchical morphology classification.

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 256×256 . 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:

6

$$\mathbf{x_norm} = (\mathbf{x} - \boldsymbol{\mu}) / \boldsymbol{\sigma} \quad (1)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ denote the channel-wise mean and standard deviation used for normalization.

10

For ImageNet-pretrained models, images were normalized using standard ImageNet statistics: $\boldsymbol{\mu} = [0.485, 0.456, 0.406]$, $\boldsymbol{\sigma} = [0.229, 0.224, 0.225]$. 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.

32

32

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 :

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

For the high-confidence Smooth vs Featured branch, the actual implementation also included YOLO11x-cls training at 448×448 resolution, where “-cls” denotes the image-classification variant. 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 thesis 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 an efficient image-level classifier, especially in case of the 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 section describes 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 3.4 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.

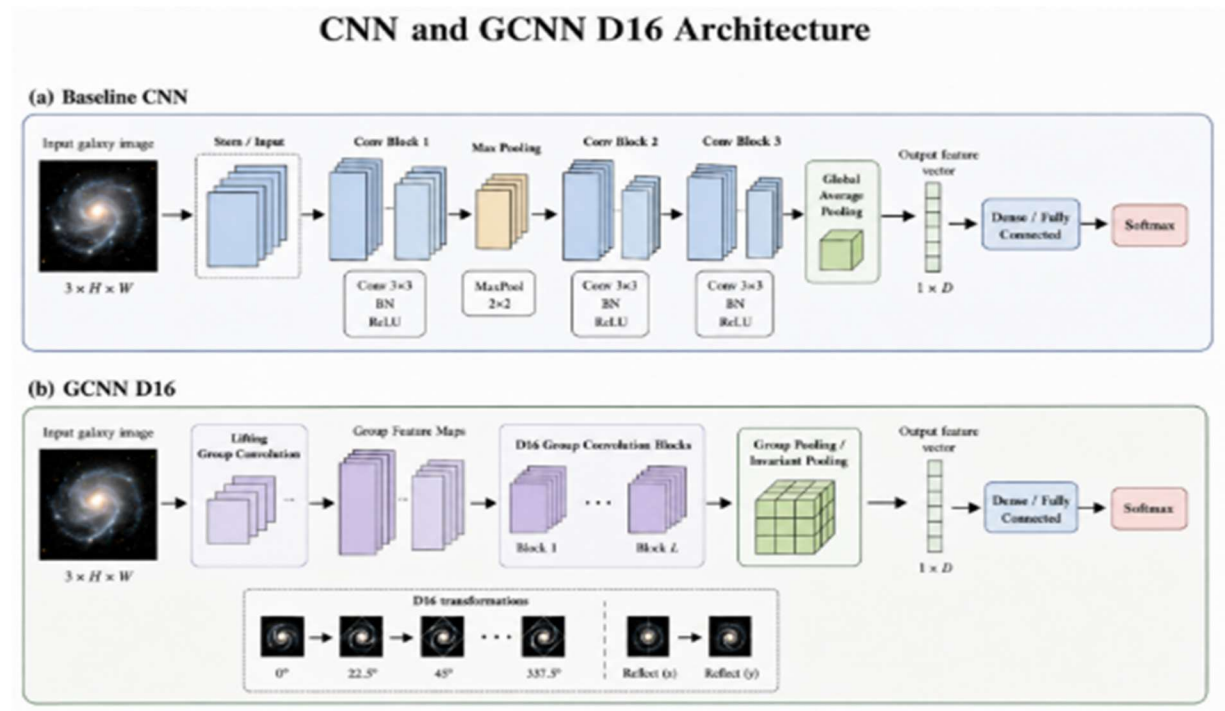


Figure 3.4 CNN family architecture used for local galaxy morphology feature extraction.

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:

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

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

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.

Figure 3.5 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.

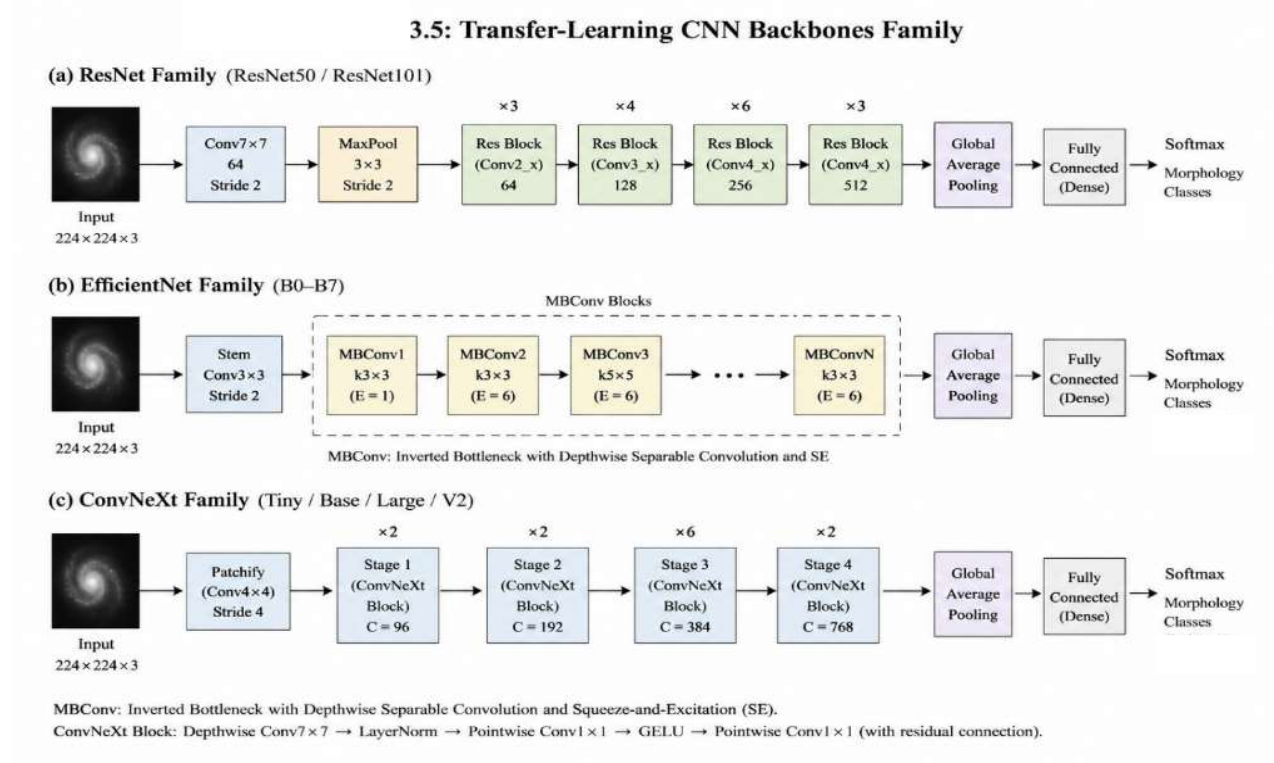


Figure 3.5 Transfer-learning CNN backbone family adapted for galaxy morphology classification.

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.

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.

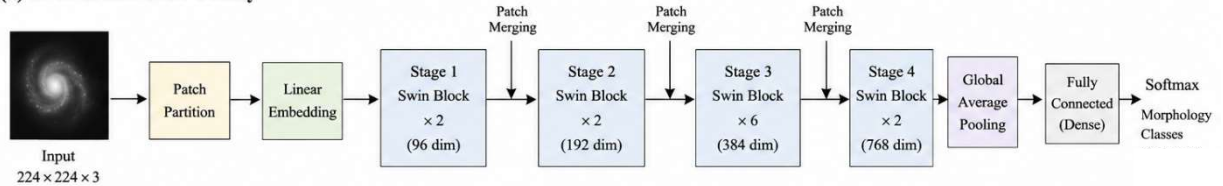
In MaxViT, the local relationships within the regions are captured using the convolutional MBConv blocks, while the attention mechanisms are used to capture the broader relationships across the entire image.

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.

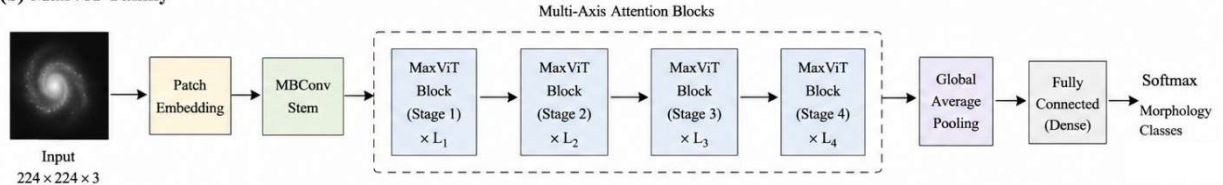
Figure 3.6 demonstrates the transformer and hybrid attention family used in this study. The representative Swin Transformer, MaxViT, and FocalNet/EVA-02 branches are followed by the feature aggregation and the task-specific classification head.

3.6: Transformer and Hybrid Attention Family

(a) Swin Transformer Family



(b) MaxViT Family



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

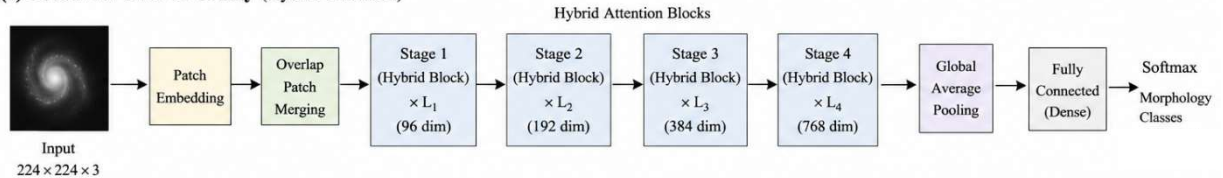


Figure 3.6 Transformer and hybrid attention architecture family used for broader spatial morphology reasoning.

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 3.7 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.

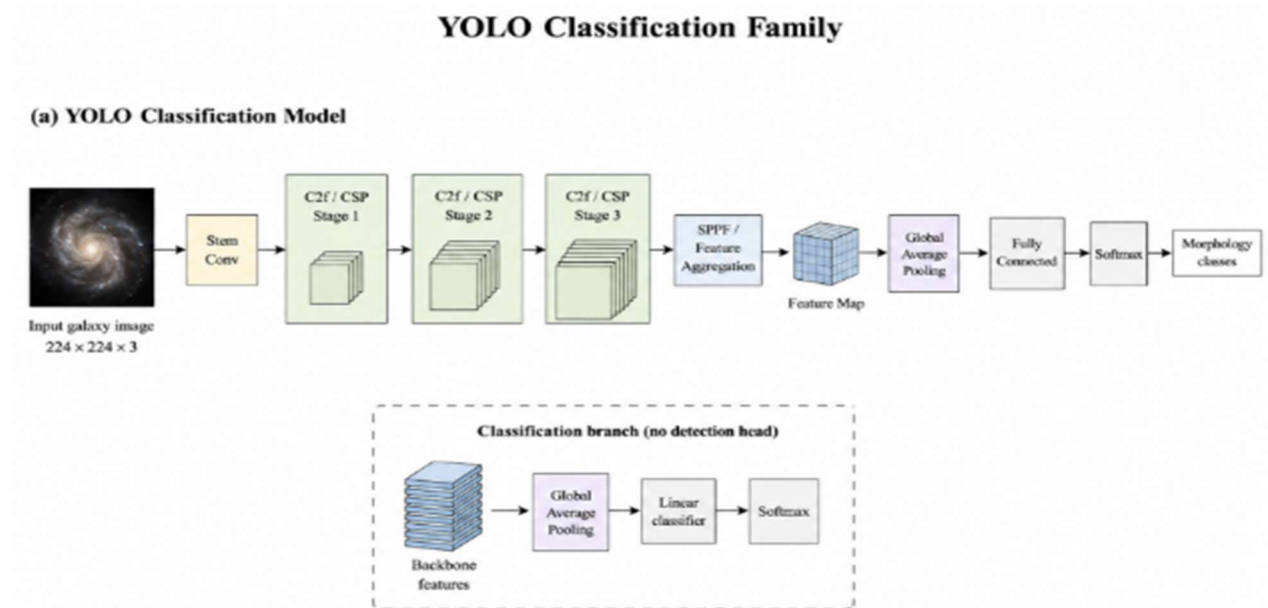


Figure 3.7 YOLO classification architecture family used as an efficient image-level morphology classifier.

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 modelweight, 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:

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

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 3.8 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.

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.

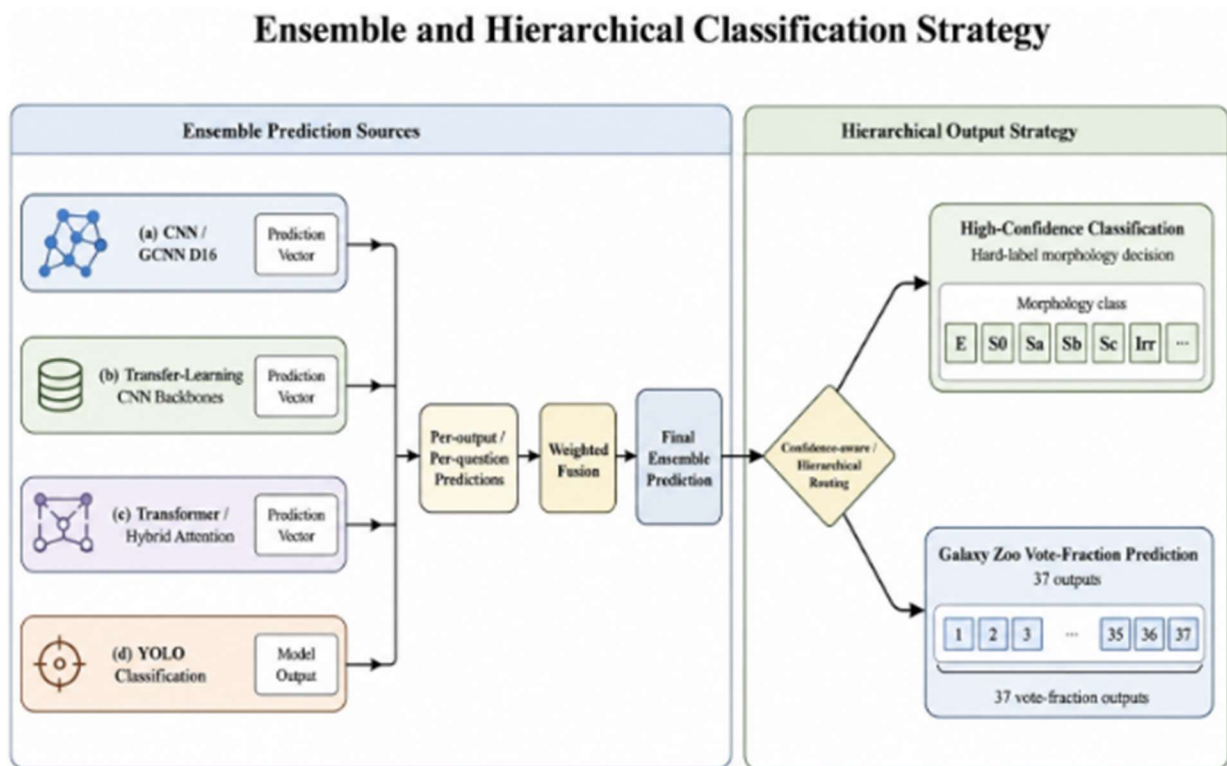


Figure 3.8 Ensemble and hierarchical prediction strategy connecting model families with 37-output Galaxy Zoo morphology prediction.

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:

14

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

9

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.

34

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.

14

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].

Chapter 4: Experimental Results and Discussion

This chapter 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 chapter 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 Chapter 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 4.1 summarizes the preprocessing outputs used for each data component in the proposed framework.

Table 4.1 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 4.1 presents representative source and preprocessed galaxy images from the Galaxy10 DECaLS and Galaxy Zoo datasets to illustrate the visual outputs used in the proposed framework.

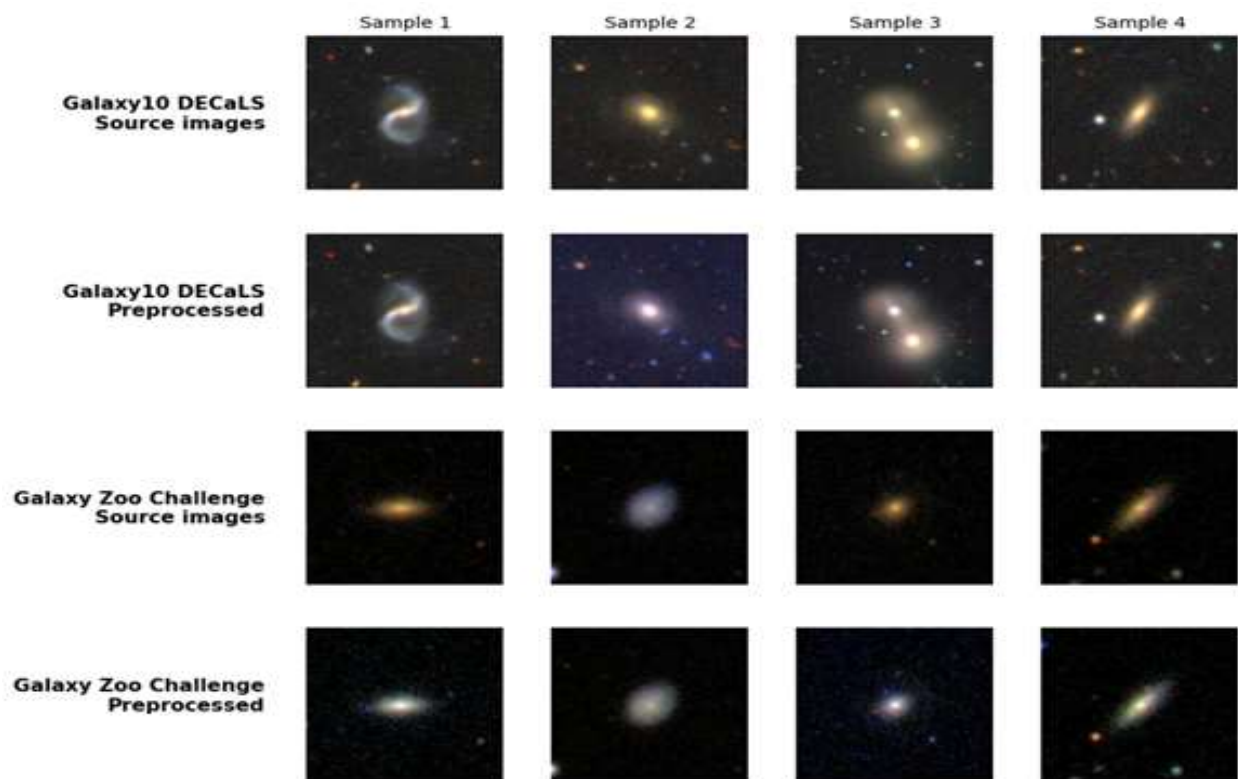


Figure 4.1 Representative source and preprocessed galaxy images from the Galaxy10 DECaLS and Galaxy Zoo datasets.

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 Chapter 3 specifies what is the methodological role of family models, whereas Chapter 4 highlights best-performing examples of each family and provides detailed analysis of the final relevant model.

Table 4.2 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.

Table 4.2 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	<i>Baseline CNN and GCNN D16</i>	<i>Galaxy10 full 10-class</i>	<i>Validation accuracy = 0.886466; Macro Recall = 0.876281;</i>	<i>Validation evidence</i>	<i>Provides the baseline and symmetry-aware reference for hard-label galaxy</i>

			Macro F1 = 0.878177		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 accuracy = 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.

4.2.1 Detailed Analysis of the ConvNeXtV2 Transfer-Learning CNN Model

A) 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 4.3 presents the per-class precision, recall, F1-score, and support values for the selected ConvNeXtV2 model on the held-out Galaxy10 DECaLS ten-class test set.

Table 4.3 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

B) Confusion matrix analysis. The confusion matrix displayed in Figure 4.2 demonstrates a strong diagonal structure, indicating that the vast majority of samples in the test set have been correctly classified. However, the remaining misclassified samples are found to belong to only a certain number of pairs of classes with similar visual appearance. Confusions are mostly found for Class 6 and Class 7, where 27 samples from Class 6 have been wrongly assigned to Class 7, while 32 samples from Class 7 have been wrongly classified as Class 6. A second case of major confusions is that between Classes 0 and 7, where 28 samples belonging to Class 0 are misclassified as Class 7, and 23 samples from Class 7 are misclassified as Class 0.

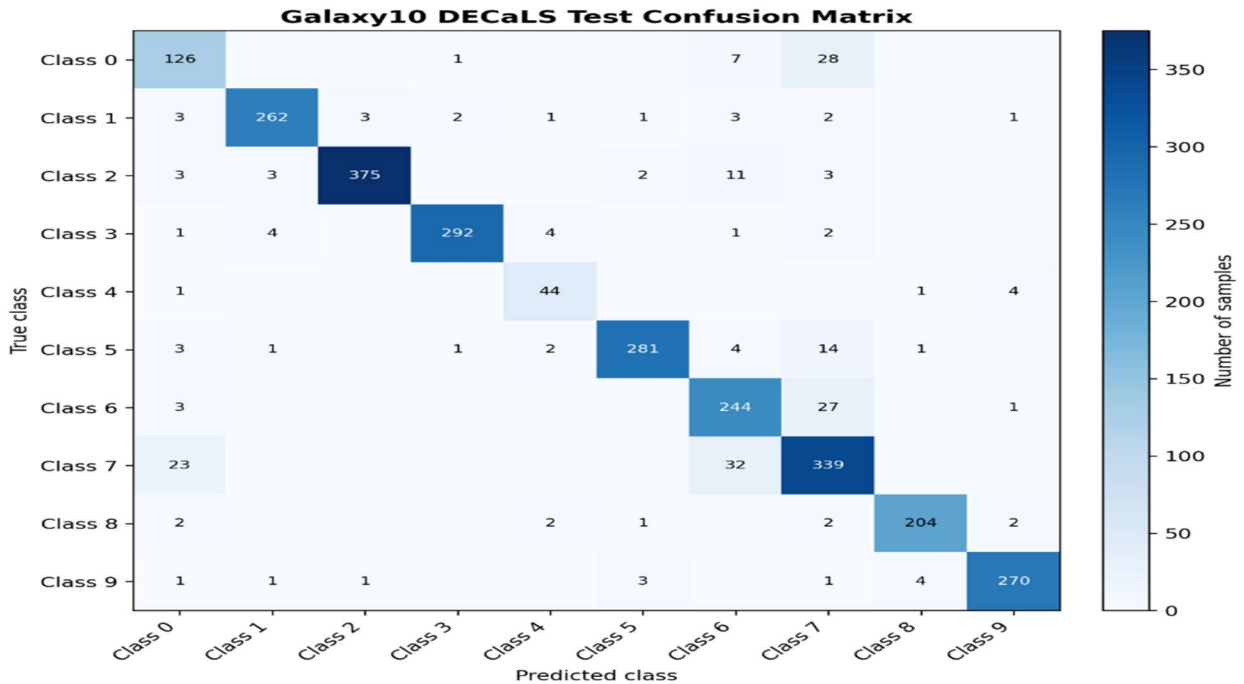


Figure 4.2 Galaxy10 DECaLS ten-class confusion matrix on the held-out test set.

C) 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 4.3 provides representative examples of correct and incorrect predictions.

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

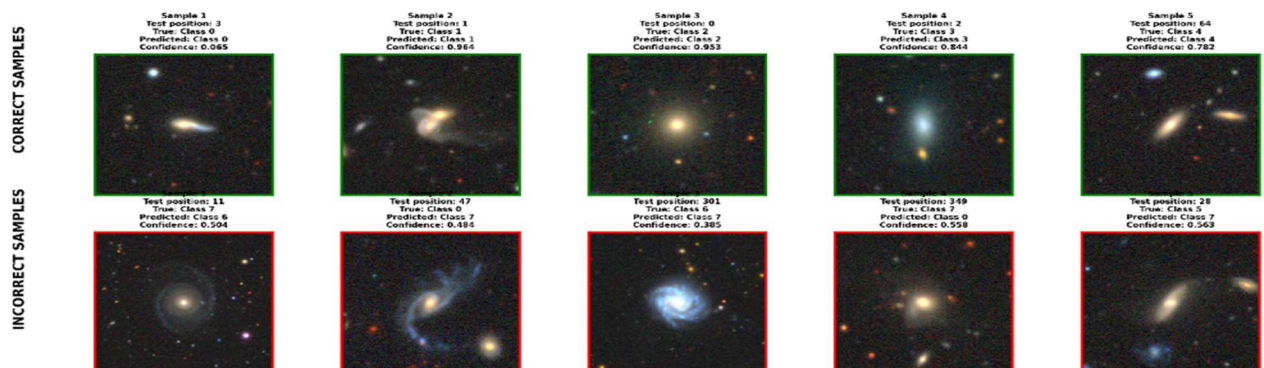


Figure 4.3 Representative correct and incorrect Galaxy10 DECaLS predictions from the ConvNeXtV2 transfer-learning CNN model.

D) 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 4.4 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).

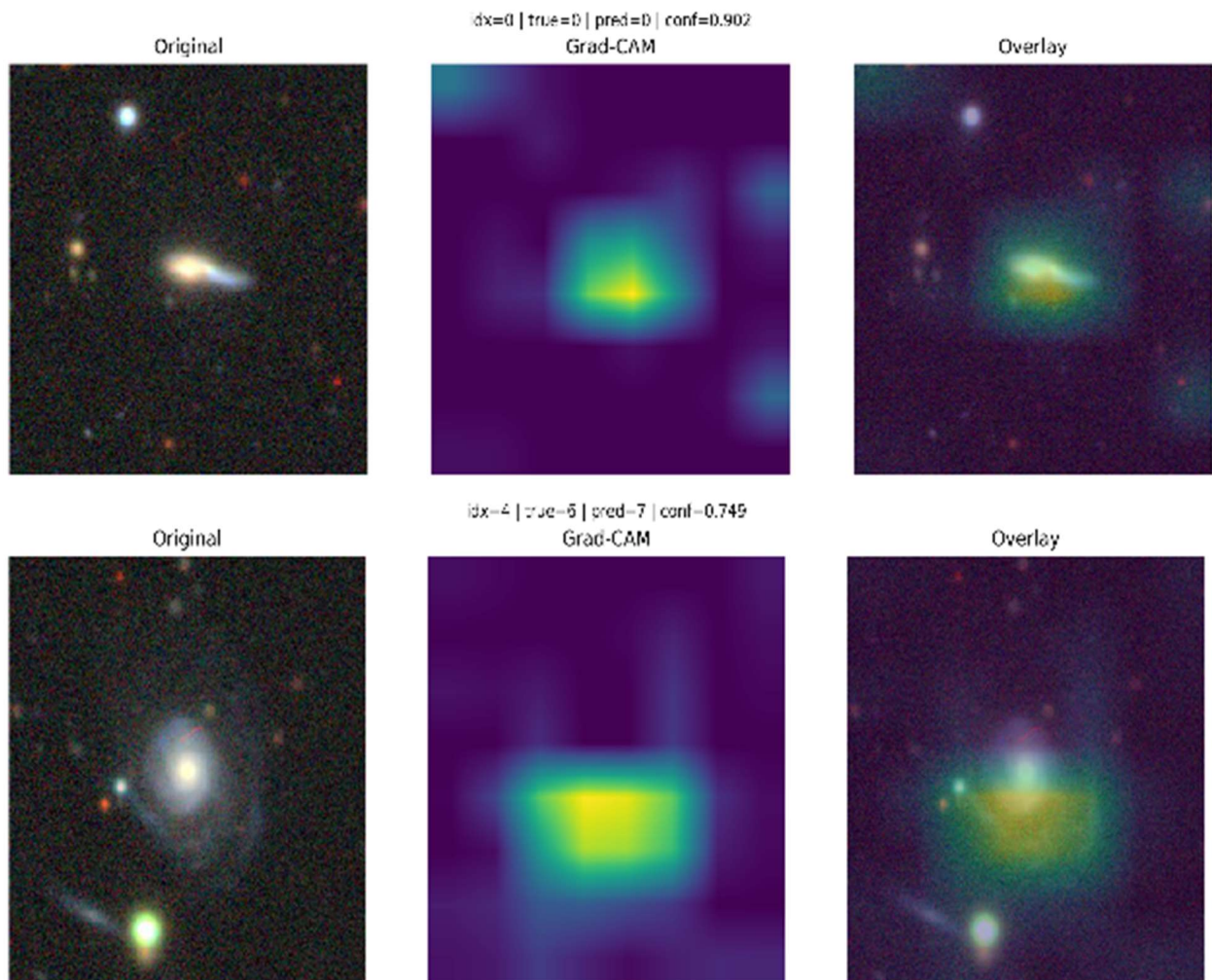


Figure 4.4 Grad-CAM visual explanations for selected Galaxy10 DECaLS predictions from the ConvNeXtV2 transfer-learning CNN model.

E) 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 4.5 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.

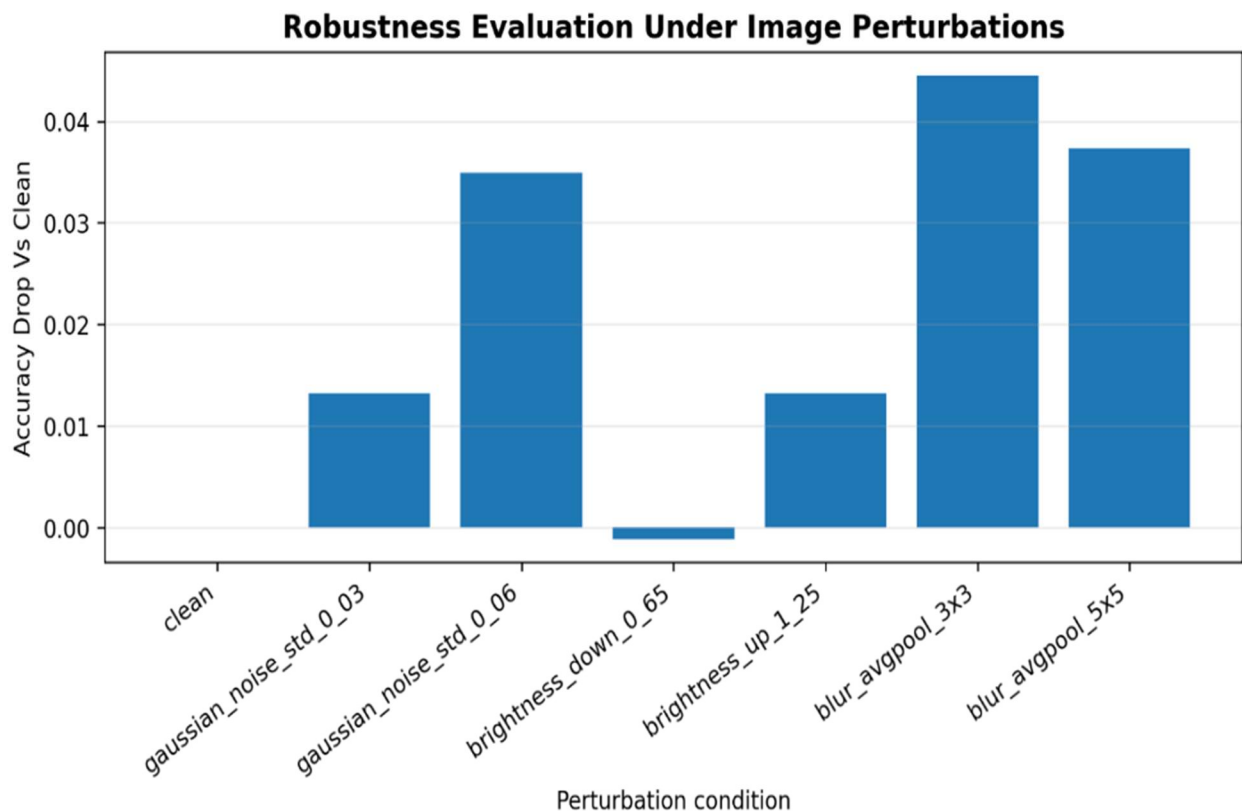


Figure 4.5 Robustness evaluation under controlled image perturbations.

F) 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 4.6 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.

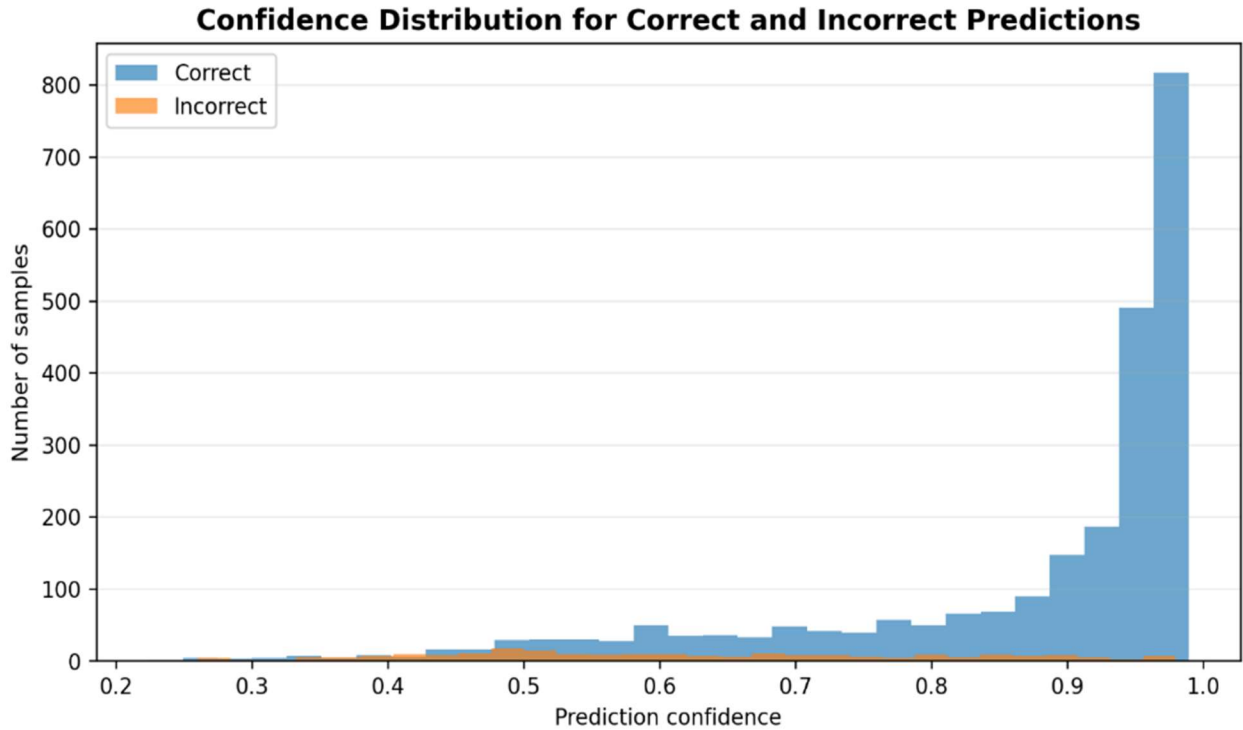


Figure 4.6 Uncertainty distribution for correct and incorrect Galaxy10 DECaLS predictions.

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 a realistic multi-class morphology classification setting.

4.3 High-Confidence Smooth vs Featured Classification

This section 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 Chapter 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 4.7 presents representative Galaxy Zoo first-gate vote-fraction examples used to construct the high-confidence Smooth vs Featured binary classification task.

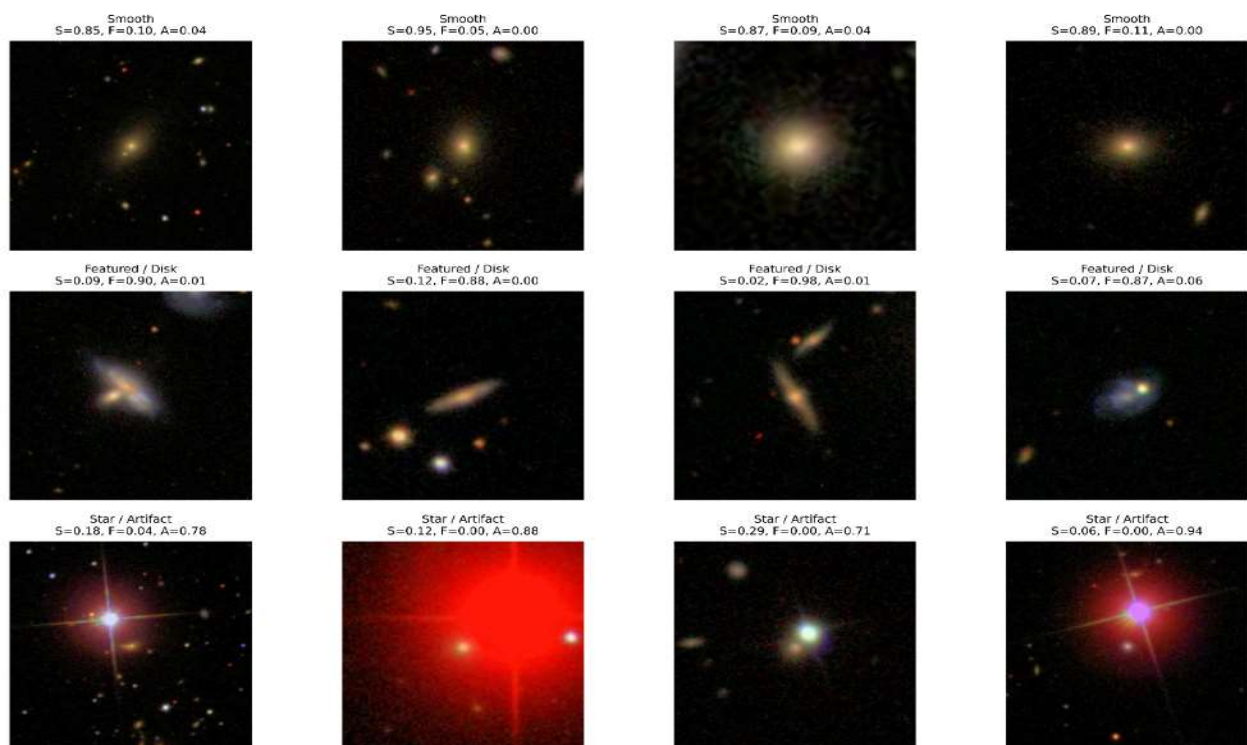


Figure 4.7 Galaxy Zoo first-gate vote-fraction examples used to construct the high-confidence Smooth vs Featured binary classification task.

A) Base Evaluation. Table 4.4 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.

The model achieved an accuracy of 99.1736% with a macro F1-score of 0.9881 on the held-out test. The Featured class got a 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.

Table 4.4 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

B) Confusion Matrix Analysis. A 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 between both binary classes.

As can be seen in Figure 4.8, 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 4.4 and showed that classification errors remained limited in both classes.

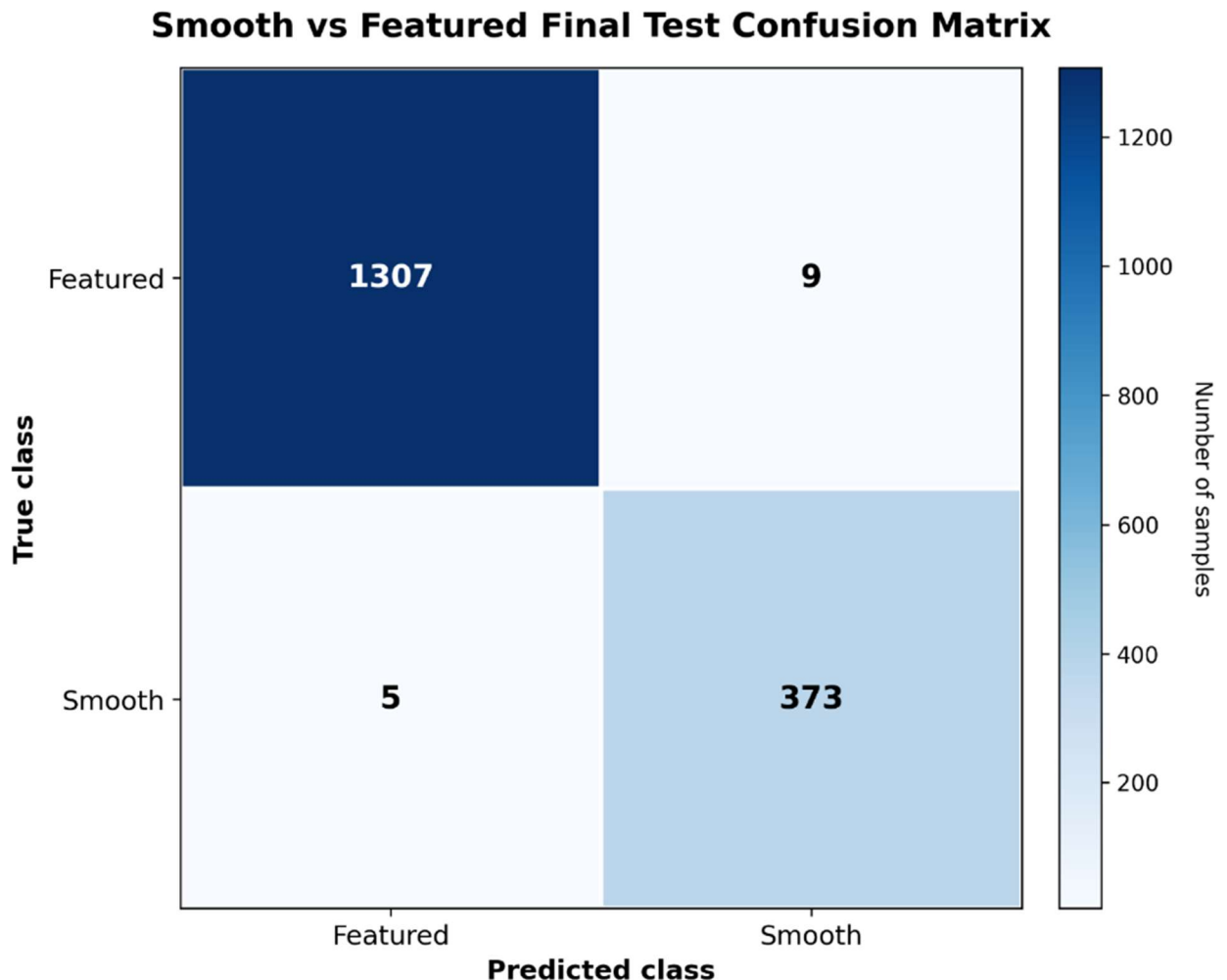


Figure 4.8 Official held-out test confusion matrix for the high-confidence Smooth vs Featured binary classification stage.

C) 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 4.9 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.



Figure 4.9 Representative correct and incorrect predictions from the high-confidence Smooth vs Featured classification stage.

D) Selective Confidence Analysis. Figure 4.10 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.

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.

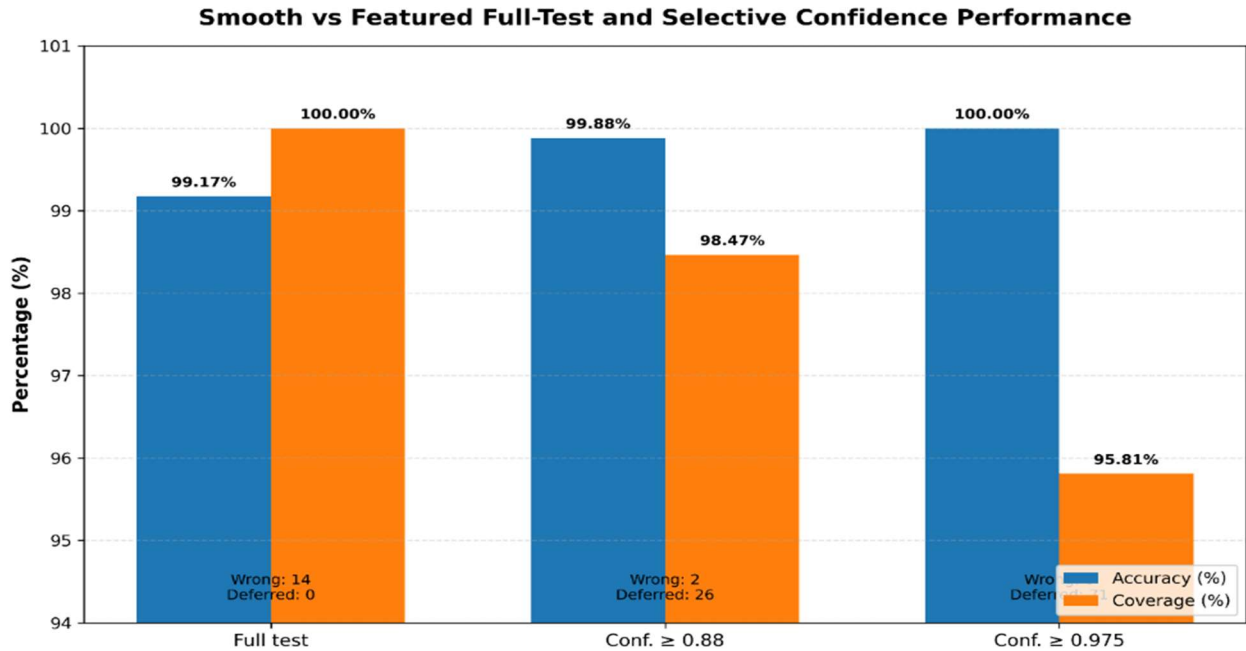


Figure 4.10 Full-test and selective-confidence performance of the Smooth vs Featured classification stage.

E) 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 4.11 compared the accuracy of the model under these conditions and the accuracy in the clean image reference condition.

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.

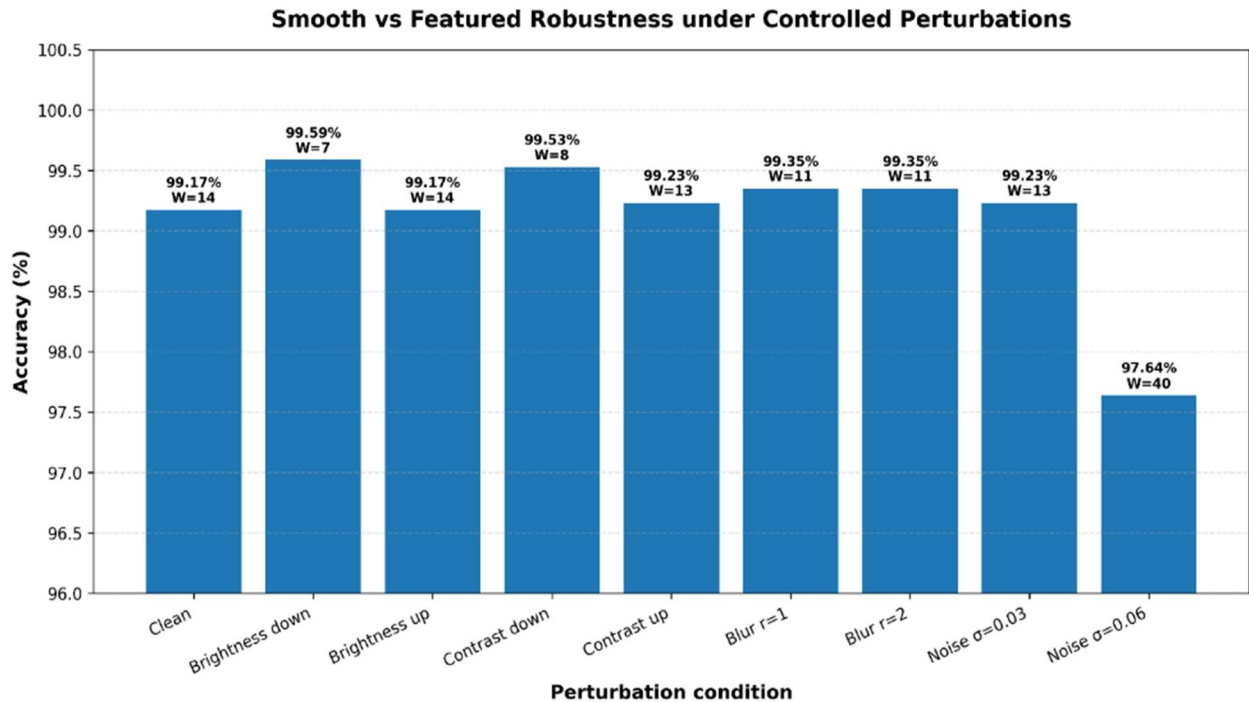


Figure 4.11 Robustness of the Smooth vs Featured classifier under controlled image perturbations.

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 section 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 Chapter 3, with ConvNeXtV2 Large used as the main held-out test architecture in the thesis

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 4.8 as an additional external benchmark for this vote-fraction branch, but it is not treated as the primary evaluation protocol of the thesis.

A) Base Evaluation. Table 4.5 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.

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.

Table 4.5 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

B) Per-output Error Analysis. Figure 4.12 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.

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.

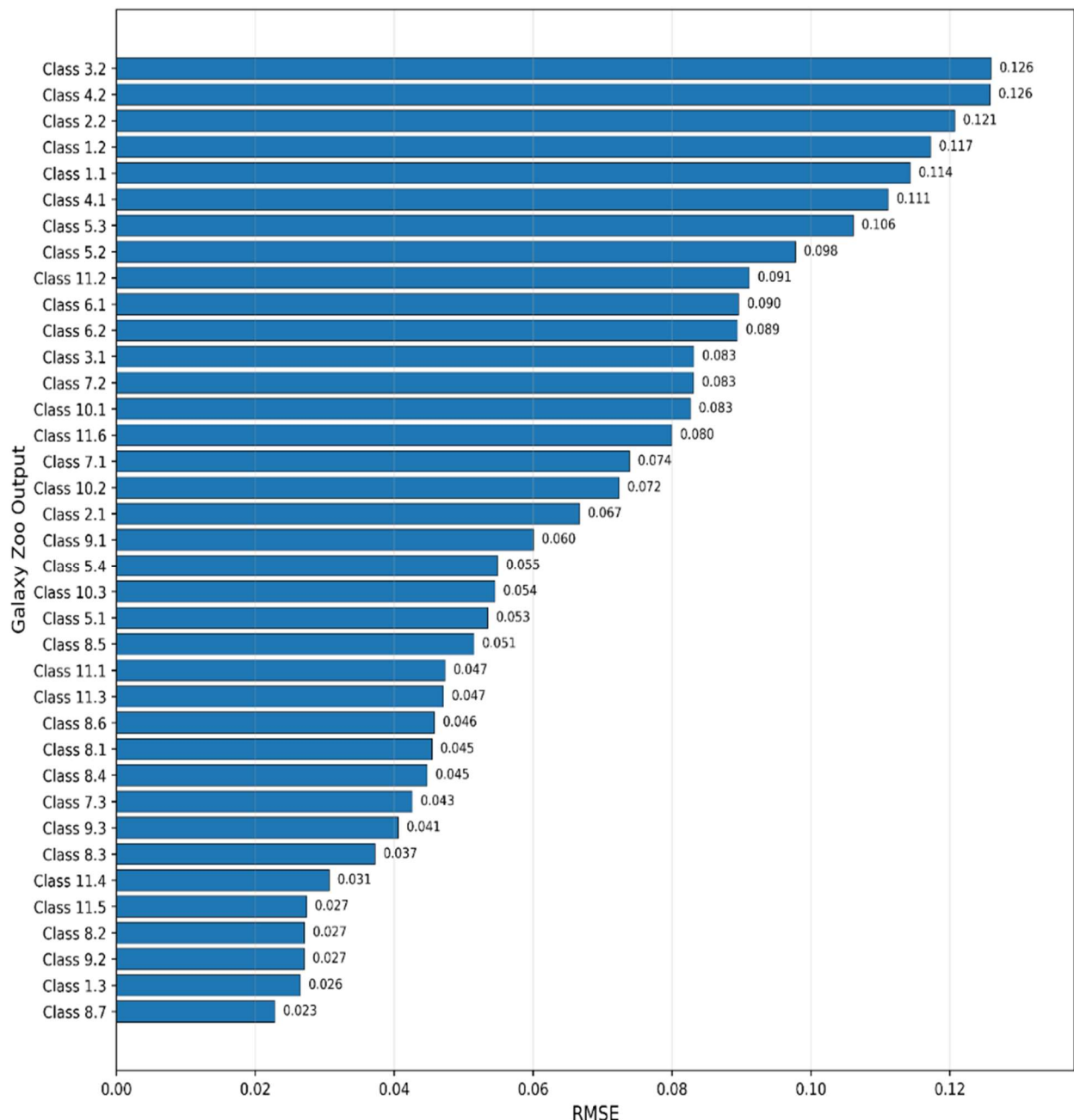


Figure 4.12 Per-output RMSE values for the Galaxy Zoo 37-output vote-fraction prediction task.

C) 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 4.13 presents the first-gate prediction analysis used to evaluate whether the 37-output vote-fraction model preserved the dominant Galaxy Zoo morphology decision.

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.

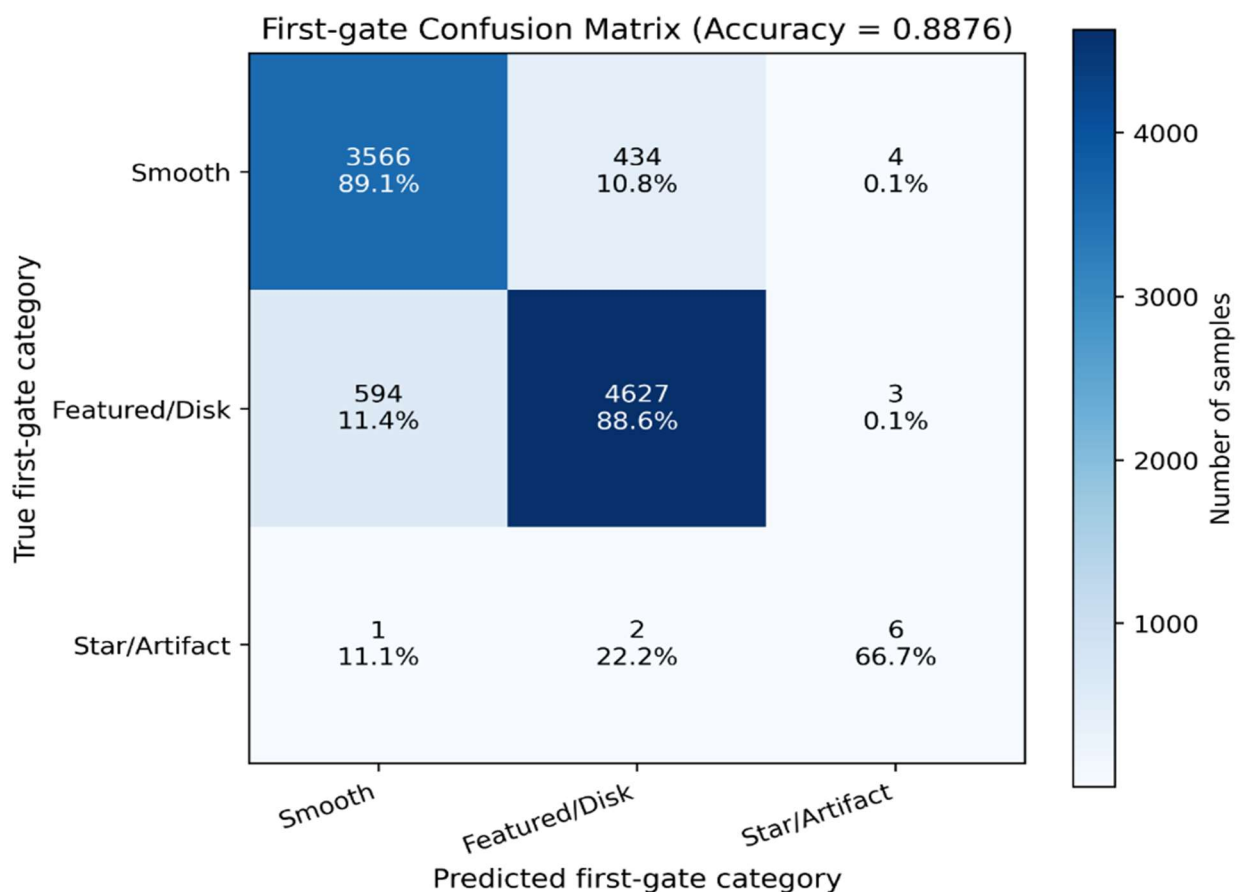


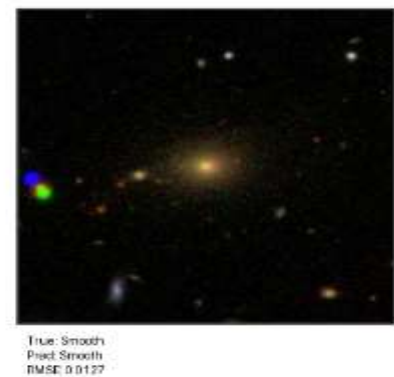
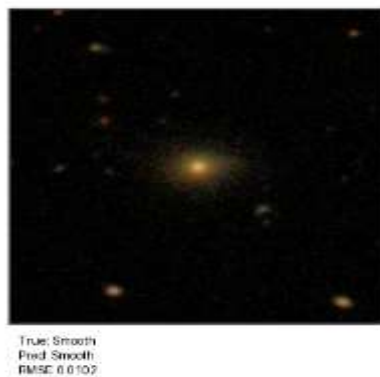
Figure 4.13 First-gate prediction analysis for the Galaxy Zoo 37-output vote-fraction model.

D) 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 ambiguous galaxy cases and prove when

the 37-output model agreed more or less with the target vote fractions. Figure 4.14 presents representative lowest-error and highest-error examples to illustrate how prediction error relates to visually clear and ambiguous galaxy morphologies.

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.

Lowest-error examples



Highest-error examples



Figure 4.14 Representative lowest-error and highest-error Galaxy Zoo vote-fraction prediction examples.

Overall, the Galaxy Zoo 37-output vote-fraction experiment proved that the framework could predict continuous morphology responses with strong agreement with 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 a component of the framework for modeling morphology uncertainty beyond single-label classification.

4.5 Reliability-Oriented Evaluation

Table 4.6 presented the main reliability evidence obtained from robustness testing, selective-confidence analysis, confidence/error diagnostics, and saliency-based visual explanations.

As can be seen from the reliability evidence in Table 4.6, 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.

Table 4.6 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.

A) 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. Figure 4.15 presents the framework-level robustness results under controlled brightness, contrast, blur, and Gaussian noise 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.

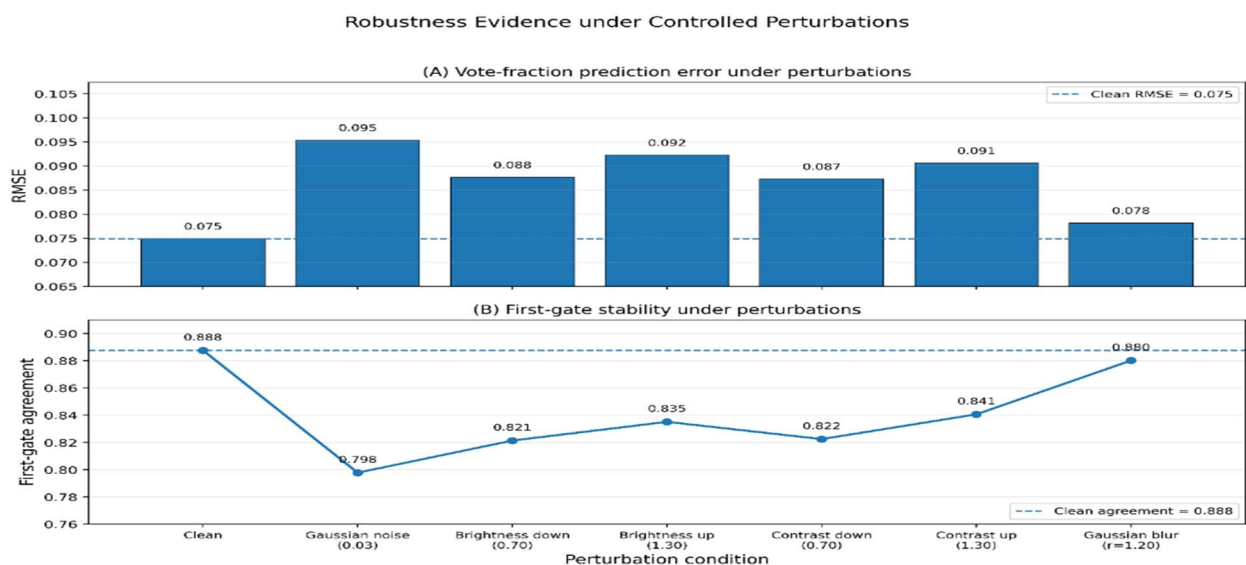


Figure 4.15 Framework-level robustness evidence under controlled image perturbations.

B) 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. Figure 4.16 presents the confidence and error evidence used to compare 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.

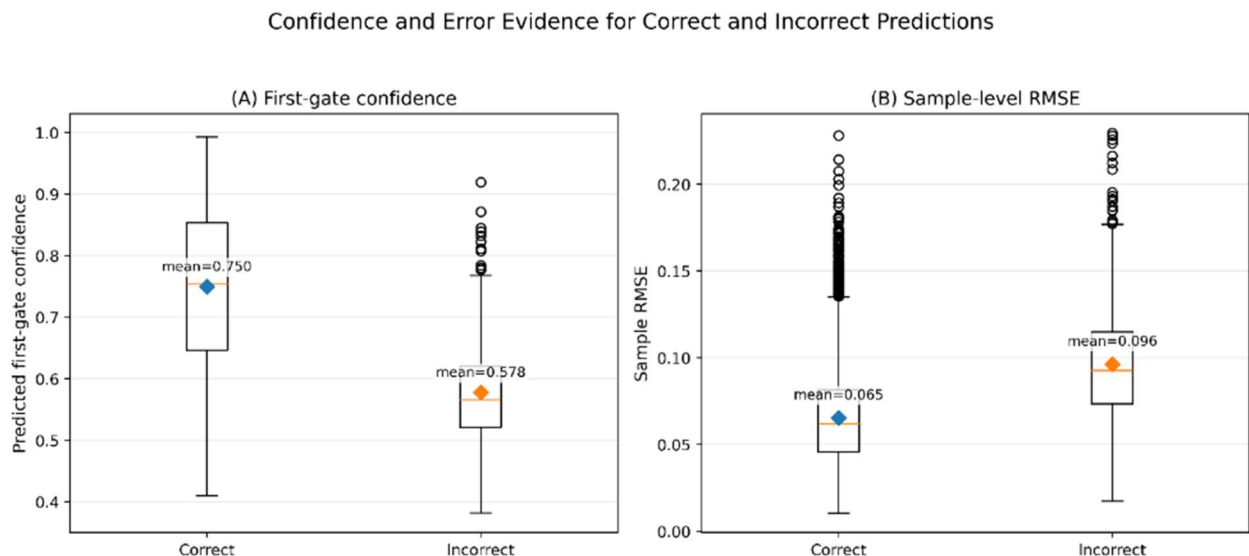


Figure 4.16 Confidence and error evidence for correct and incorrect galaxy morphology predictions.

C) 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 4.17 presents 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.

Lowest-error explanation examples

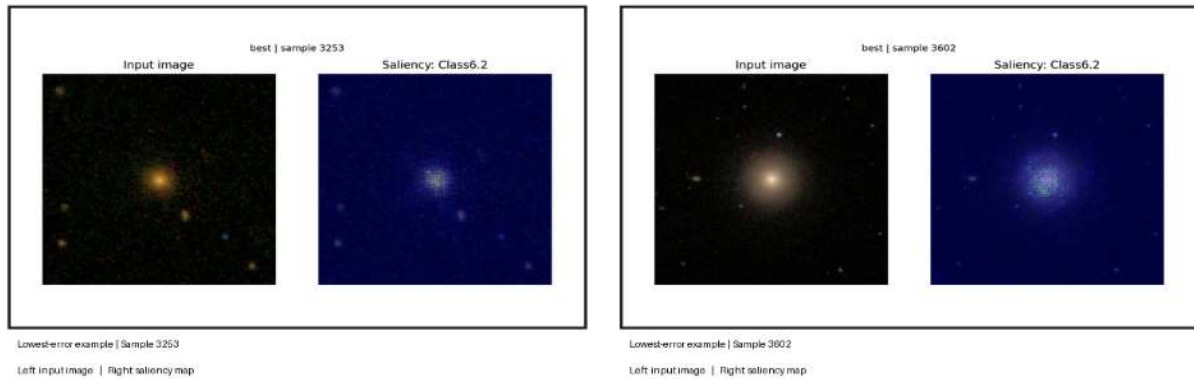


Figure 4.17 Saliency-based visual explanation evidence for selected lowest-error and highest-error Galaxy Zoo predictions.

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 Chapter 4 and places the proposed framework 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.

According to Table 4.7, the highest reported accuracy for Galaxy10 DECaLS hard-label classification among the selected related studies is about 97.97%, obtained by Fernando et al. [11] using VGG19 transfer learning. In this thesis, the Galaxy10 DECaLS hard-label branch achieved 91.58% accuracy and a macro F1-score of 0.9073 using the ConvNeXtV2 transfer-learning CNN model. While this is not the highest reported Galaxy10 DECaLS accuracy among the selected studies, it constitutes the main Galaxy10 DECaLS hard-label classification result of the proposed framework and is confirmed by class-level error analysis, robustness testing, confidence behaviour, and visual explanation evidence.

Fernando et al. [11] obtained the reported 97.97% by calculating accuracy separately for each of the ten classes against the remaining nine classes and then averaging the ten class-wise results. In this calculation, images from the other classes that are correctly rejected are counted as true negatives, which increases the class-wise accuracy and makes it different from our direct sample-level ten-class accuracy of 91.58%. A more balanced comparison is therefore provided by macro F1-score, where Fernando et al. reported 89.68%, while our held-out test achieved 90.73%. The remaining errors in our model were mainly concentrated in the visually overlapping Classes 0, 6, and 7.

Table 4.7 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.
Mao et al. [16]	Galaxy10 DECaLS and GZD-5 classification	DenseNet-SE4 with attention	88.96% accuracy	Attention-enhanced comparison for visually similar galaxy classes.
Chen et al. [18]	Galaxy10 DECaLS classification	Semi-supervised learning with uncertainty estimation	76.8% accuracy with 2,000 labeled samples	Uncertainty-aware reference under limited-label learning.

Fernando et al. [11]	Galaxy10 DECaLS classification	VGG19 Transfer Learning	97.97% accuracy & Macro F1= 0.8968	Uses class-wise one-versus-rest accuracy, so the reported accuracy is not directly comparable with this framework's sample-level ten-class accuracy; the macro F1 comparison slightly favours this framework.
This Framework: 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 4.8 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 the 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.

Table 4.8 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 37-output vote-fraction prediction	Rotation-Invariant CNN with Model Averaging	Private RMSE = 0.07492; >99% accuracy for most high-agreement questions	Direct published benchmark; this framework achieved a lower Private RMSE of 0.07440

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. [17]	Galaxy Zoo morphology questions under cross-survey shift	Unsupervised domain adaptation for Galaxy Zoo morphology prediction	79.5%–96.7% accuracy across target-domain morphology questions; 87.5% for Smooth or Featured	Supports the importance of robustness and dataset-shift evaluation for first-level morphology prediction.
This Framework: Smooth vs Featured stage	Galaxy Zoo high-confidence binary morphology routing	YOLO11x classification family	<u>99.1736%</u> accuracy; macro F1 = 0.9881	Shows strong first-level morphology routing using high-confidence vote-fraction labels.
This Framework: 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 Framework: 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.

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 complementary information that helps to interpret the results.

Chapter 5: Conclusion and Future Work

60 This thesis describes and evaluates a structured deep learning approach for galaxy morphology classification and analysis. In particular, the approach strives to transcend the one-model, one-metric paradigm by combining three different tasks: (1) ten-class hard label classification of Galaxy10 DECaLS, (2) high-confidence Smooth-versus-Featured routing inspired by Galaxy Zoo voting fractions, and (3) continuous response prediction of 37 Galaxy Zoo vote-fraction outputs. This combination is supported by the set of reliability-oriented analyses, including robustness testing, confidence analysis, Grad-CAM, and saliency mapping. Thus, the framework combines morphological classification with voting fraction prediction while maintaining task-specific evaluation protocols.

The empirical results indicate that the described approach shows high predictive power in all settings under consideration. For the Galaxy10 DECaLS ten-class task, the ConvNeXtV2 transfer learning model reached 91.58% accuracy and 0.9073 macro F1-score. In the high-confidence Smooth-versus-Featured routing setting, YOLO11x reached 99.1736% accuracy and 0.9881 macro F1-score, thus showing reliable first-level morphology routing when the target label is formed based on strong agreement of the volunteer vote fractions. For the 37-response Galaxy Zoo task, ConvNeXtV2 Large reached 0.074912 RMSE and 0.042315 MAE. In a post-deadline Kaggle submission, 0.07433 public RMSE and 0.07440 private RMSE were obtained. This submission serves as an external benchmark, while the held-out test dataset serves as the primary basis of scientific evaluation.

79 Thus, the described findings answer the main research question within the considered datasets and experimental settings. Deep learning helps to augment galaxy morphology classification and analysis provided that the predictive accuracy is supplemented with confidence-based assessment, continuous voting fraction representation, robustness evaluation, and visual explanations. The main contribution of this thesis is not the performance of a single architecture. Instead, it consists in the combination of model-family comparison, hierarchical morphology routing, continuous response prediction, and reliability-oriented evaluation within a single coherent framework. From an astronomical point of view, this framework can help to conduct large-scale morphology studies by providing initial predictions, detecting high-confidence examples, and selecting the ambiguous ones that require further

24

20

25

expert analysis. Therefore, the approach should be understood as a decision support approach rather than a replacement for the scientific interpretation of images.

Furthermore, the results show that accuracy is not sufficient for evaluation of scientific image-analysis systems. Perturbation experiments show how the performance of the system changes under modified images, while confidence analyses help to distinguish many correct from incorrect predictions. Grad-CAM and saliency maps show that model attention often matches morphologically meaningful regions. However, this explanation should be understood as evidence supporting the predictions rather than proof of learning causal or physically complete representations. Taken together, the reliability analyses provide the broader basis for the understanding of model behavior, but they do not obviate the need for astronomical validation.

Several limitations define the scope of the conclusions. The main experiments have been performed on Galaxy10 DECaLS and Galaxy Zoo-related datasets, whose results cannot be generalized to other surveys, instruments, redshift ranges, image resolutions, or observational conditions. Some categories are always ambiguous or visually overlapping, especially when structural details are low. Furthermore, the robustness analysis included a controlled set of perturbations rather than the whole range of survey effects, and the visual explanation assessment was mostly qualitative. The post-deadline Kaggle scores should also be understood as additional benchmark evidence, but not a formal final leaderboard entry.

82

Future work should focus on cross-survey validation and adaptation of the approach to data from different astronomical surveys. Advanced uncertainty estimation and confidence calibration methods, including probabilistic and Bayesian techniques, could improve the detection of ambiguous cases and help human-in-the-loop analysis. Robustness evaluation could be expanded to cover realistic variations in point spread function, background noise, resolution change, and specific survey effects. Explainability should be quantified by comparing the highlighted regions with well-known morphological features like spiral arms, bars, bulges, and edge-on disks. Finally, the framework could be developed as an AI-assisted analysis tool reporting the class predictions, voting fraction estimates, confidence measures, and explanations.

62

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., Zhang, S., Fu, L., Xiao, H., & Shu, C. “Galaxy Morphology Classification via Deep Semi-supervised Learning with Limited Labeled Data”. *The Astrophysical Journal Supplement Series* 279(1), Article 17, 2025. DOI:10.3847/1538-4365/addb4c

- [9] [Yao, Y., et al., 2024] Yao, Y., Zhang, J., Du, P., & Dong, S. “A Galaxy Image Augmentation Method Based on Few-shot Learning and generative adversarial networks”. *Research in Astronomy and Astrophysics* 24(3), Article 035015, 2024. DOI:10.1088/1674-4527/ad2128
- [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] [Fernando et al., 2024] Fernando, T., Rathnayake, S., & Dissanayaka, K. “Galaxy Morphology Classification Based on VGG19 Deep Convolutional Neural Network”. *6th International Conference on Advancements in Computing (ICAC)*, pp. 229–234, 2024. DOI:10.1109/ICAC64487.2024.10851157
- [12] [Parker et al., 2024] Parker, L., Lanusse, F., Golkar, S., et al. “AstroCLIP: a cross-modal foundation model for galaxies”. *Monthly Notices of the Royal Astronomical Society* 531(4), pp. 4990–5011, 2024. DOI:10.1093/mnras/stae1450
- [13] [Cao, J., et al., 2024] Cao, J., Xu, T., Deng, Y., Deng, L., Yang, M., Liu, Z., & Zhou, W. “Galaxy morphology classification based on Convolutional Vision Transformer (CvT)”. *Astronomy & Astrophysics* 683, A42, 2024. [DOI:10.1051/0004-6361/202348544](https://doi.org/10.1051/0004-6361/202348544)
- [14] [Urechiatu, R., & Frincu, M., 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., Cortesi, A., Ribeiro, U., Dias, L. O., et al. “An extended catalogue of galaxy morphology using deep learning in Southern Photometric Local Universe Survey Data Release 3”. *Monthly Notices of the Royal Astronomical Society* 528(3), pp. 4188–4208, 2024. DOI:10.1093/mnras/stad3956
- [16] [Mao et al., 2025] Mao, Y., Tu, L., Xu, Z., Jiang, Y., & Zheng, M. “Galaxy morphology classification based on DenseNet-SE4 algorithm”. *Research in Astronomy and Astrophysics*, 25(8), 085009, 2025. DOI:10.1088/1674-4527/ade22c
- [17] [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 adaption”. *Monthly Notices of the Royal Astronomical Society* 537(2), pp. 640–649, 2025. DOI:10.1093/mnras/staf025

[18] [Chen et al., 2025] Chen, Z., Xu, Z., & Chen, W. “Semi-supervised learning for galaxy morphology classification based on diversity enhancement and uncertainty estimation”. *Research in Astronomy and Astrophysics*, 25(10), 104005, 2025. DOI:10.1088/1674-4527/adf712

[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] [NASA APOD, 2017] NASA Astronomy Picture of the Day. “UGC 1810: Wildly Interacting Galaxy from Hubble”. May 10, 2017. Image credit: NASA, ESA, Hubble, and HLA; processing and copyright: Domingo Pestana. Available at: <https://apod.nasa.gov/apod/ap170510.html> (Accessed: July 22, 2026).