



MASTER'S THESIS PRESENTATION

Deep Learning Framework for Galaxy Morphology Classification and Analysis

Prepared by:

Ahmad Yousef Alkahlout

Mohammed Taleb Ragash Mohammed

Supervisor: Dr. Mohamed Eassa

College of Informatics · Artificial Intelligence · Midocean University

MSc in Artificial Intelligence

ABSTRACT

Study Summary

Problem, framework, datasets, tasks, evaluation, and final results.

Problem

The increasing number of imaged galaxies makes manual morphology classification impossible for large-scale research, while many studies evaluate only predictive accuracy and neglect robustness, uncertainty, and interpretability.

Framework

This thesis proposes a confidence-aware deep learning framework and evaluates CNNs, GCNNs, transfer-learning CNNs, transformer-based and hybrid models, YOLO classifiers, and ensembling approaches.

91.58%

Galaxy10 held-out accuracy

0.9073

Galaxy10 Macro F1

99.17%

Smooth/Featured accuracy

0.074912

Galaxy Zoo held-out RMSE

0.042315

Galaxy Zoo held-out MAE

Data and tasks

Galaxy10 DECaLS was used for ten-class hard-label classification. Galaxy Zoo Challenge was used for high-confidence Smooth-versus-Featured classification and 37 continuous vote-fraction outputs.

Evaluation evidence

Accuracy, precision, recall, Macro F1, RMSE, MAE, controlled perturbations, confidence analysis, Grad-CAM, and saliency maps.



Key Point — The thesis is not one model experiment; it is a structured framework linking three prediction tasks with reliability-oriented evaluation.



CHAPTER 1 · SECTION 1.1

Overview

Why galaxy morphology classification became an AI problem.

Scientific context

Galaxy morphology classification is important because the visible structure of a galaxy carries information about its origin and evolution.

Scale problem

Modern surveys such as SDSS and DECaLS increased the number of available galaxy images, making manual classification difficult and time-consuming.

AI motivation

Deep learning can learn visual features directly from images, but scientific use also requires robustness, uncertainty awareness, and interpretability.

Why this study starts here

- Large visual diversity between galaxies
- Ambiguous boundaries between some morphology classes
- Need for automated classification with reliability evidence

❏ **Key Point** — The application is astronomy, but the AI problem is large-scale visual classification with reliability, confidence, and interpretability requirements.

Problem Definition

The gap between accurate prediction and reliable scientific classification.

Interpretability

A correct label is not enough if the model may rely on irrelevant image regions rather than spiral arms, bars, or the central bulge.

Robustness

Brightness, contrast, blur, and noise can alter image appearance. Scientific use requires understanding how predictions change.

Uncertainty

Some galaxies have weak or mixed morphology. A single final label may hide ambiguity and create false confidence.

Unified framework gap

Existing solutions often study accuracy, robustness, uncertainty, or explainability separately rather than integrating them.

What deep learning framework can provide accurate galaxy morphology classification while remaining robust, interpretable, and uncertainty-aware?



Key Point — The research question is not "Which model has the highest accuracy?" but "How can prediction and reliability evidence be integrated?"

Thesis Objectives

Six objectives aligned with the proposed framework.

01	02	03
Create baseline CNN models and assess direct morphology feature learning.	Apply transfer learning and pretrained backbones to improve generalisation.	Move from controlled hard labels toward confidence-based vote-fraction representations.
04	05	06
Evaluate classification using accuracy and F1-score and regression using RMSE and MAE.	Test robustness and confidence under controlled image perturbations.	Use Grad-CAM and saliency maps to inspect whether decisions focus on relevant structures.

 **Key Point** — Every objective maps to an evaluated component; none is stated without corresponding experimental evidence in Chapters 3 and 4.

Thesis Contribution

What the study contributes without claiming a new architecture.



Structured model-family evaluation

Systematic comparison of baseline CNNs, GCNNs, pretrained CNNs, transformers, YOLO classifiers, and fusion strategies.



Robustness evaluation

Controlled image perturbations used to measure stability and sensitivity.



Explainable AI evidence

Grad-CAM and saliency maps used to inspect image regions influencing predictions.



Confidence-aware extension

Progression from one hard label to high-confidence routing and continuous vote-fraction prediction.



Uncertainty evidence

Confidence and error behaviour used to identify ambiguous or lower-reliability predictions.



Key Point — The contribution is methodological integration and disciplined evaluation—not the invention of a new neural-network layer.

Thesis Organization

The presentation follows the same order as the thesis.

1

1 · Introduction

Problem, objectives, contribution

2

2 · Background & Related Work

Concepts, prior studies, research gap

3

3 · Proposed Methodology

Datasets, preprocessing, models, evaluation

4

4 · Results & Discussion

Performance, errors, reliability, comparison

5

5 · Conclusion & Future Work

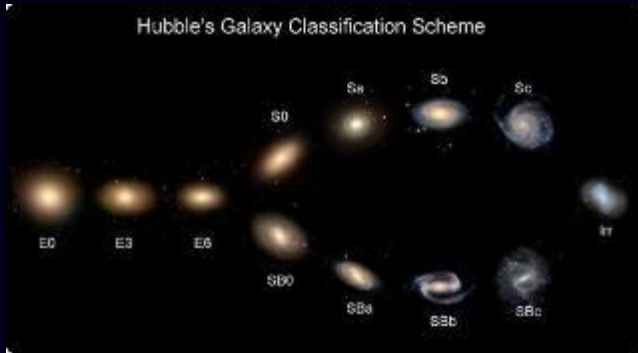
Findings, limitations, next steps



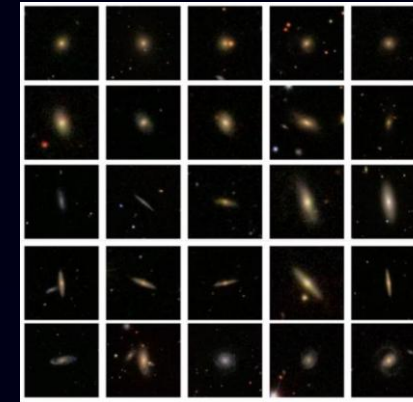
Key Point — The slide sequence mirrors the thesis, which makes it easy for the committee to connect each claim to its documented chapter.

From Visual Morphology to Large-Scale Classification

Why automated morphology analysis became necessary.



Hubble morphological classification scheme.



Examples showing high visual diversity among morphology classes.

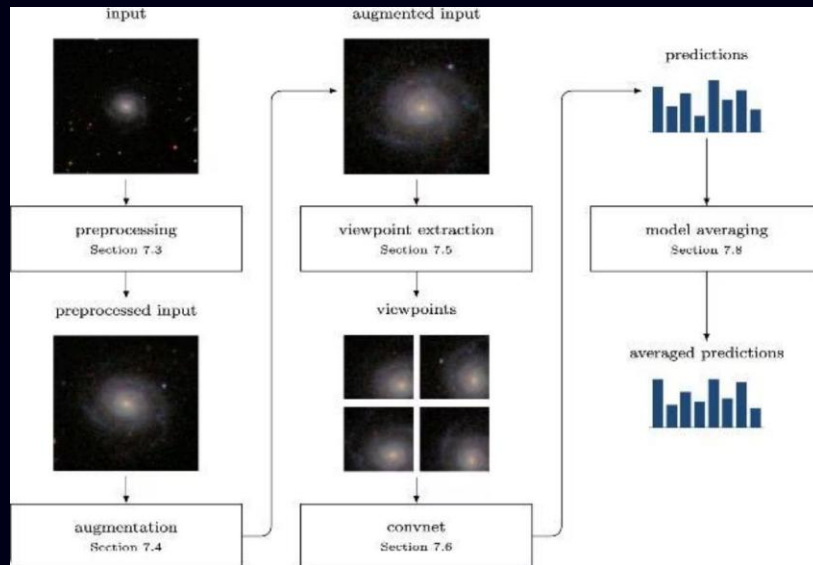
Background logic

- Hubble and de Vaucouleurs schemes organised galaxies by visible morphology.
- Manual visual inspection became impractical as SDSS and DECaLS increased image volume.
- Galaxy Zoo introduced citizen-science labels and vote fractions that preserve classifier agreement.
- Visual overlap between classes makes the task difficult for humans and computer-vision models.

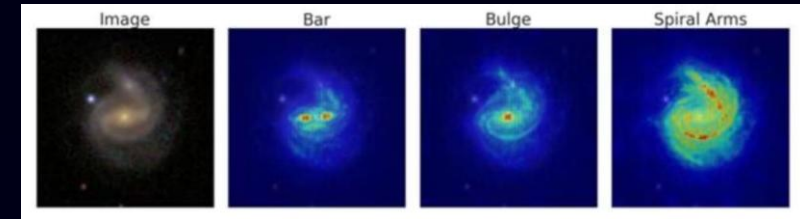
☐ **Key Point** — The value of Galaxy Zoo is not only more labels; vote fractions preserve agreement and ambiguity that a single hard label hides.

Deep Learning and Explainability Background

Local feature learning, global structure, invariance, and visual inspection.



Rotation-aware deep learning pipeline adapted from Dieleman et al.



Example saliency map highlighting discriminative regions.

Deep learning role

CNNs learn image features directly; transfer learning reuses pretrained representations; transformers model broader spatial relationships; semi-supervised methods reduce label dependence.



Key Point — Grad-CAM and saliency support inspection of model focus, but they are evidence for interpretation—not proof of causal astronomical reasoning.

Probabilistic Labels and Rotation Invariance

Related studies summarized by method, dataset, advantage, and limitation.

Walmsley et al. [1]

Method / Dataset Bayesian deep learning (Zoobot); Galaxy Zoo / DECaLS; about 300,000 galaxies.

Advantage Predicts distributions of volunteer answers; supports the move beyond one hard label.

Limitation Volunteer-generated answers remain the main data limitation.

Dieleman et al. [2]

Method / Dataset Rotation-invariant CNN with model averaging; Galaxy Zoo 2; 37 vote-fraction outputs.

Advantage Private RMSE = 0.07492; strong high-agreement question performance.

Limitation Prediction averaging increases computational cost.



Key Point — These studies support the use of vote fractions and orientation-aware reasoning rather than relying only on one final class label.

Explainability and CNN Benchmarks

Related studies summarised by method, dataset, advantage, and limitation.

Bhambra et al. [3]

Method / Dataset Grad-CAM and saliency maps for explainable galaxy classification.

Advantage Shows networks can focus on spiral arms, cores, and asymmetry.

Limitation Provides explainability evidence rather than a new benchmark.

Domínguez Sánchez et al. [7]

Method / Dataset Deep CNN on SDSS morphology classification.

Advantage About 95% success rate and direct feature learning from pixels.

Limitation Different dataset and not an integrated reliability framework.



Key Point — These studies link directly to our use of visual explanation evidence and CNN baselines, while showing that high accuracy alone is not the full research gap.

Limited Labels and Augmentation

Related studies summarised by method, dataset, advantage, and limitation.

Luo et al. [8]

Method / Dataset Semi-supervised CNN and GAN-based learning on Galaxy10 DECaLS.

Advantage About 84% accuracy while reducing dependence on fully labelled data.

Limitation The approach was reported as unstable.

Yao et al. [9]

Method / Dataset FSL-GAN data augmentation with AlexNet, VGG, and ResNet on Galaxy10 DECaLS.

Advantage Shows augmentation effects differ by class and model.

Limitation No one-size-fits-all augmentation method for all morphology classes.



Key Point — These studies are linked by Galaxy10 DECaLS and data-efficiency. Their relevance is that augmentation and semi-supervised learning must be tested, not assumed to solve all classes.

Transfer Learning on Galaxy10 DECaLS

Related studies summarised by method, dataset, advantage, and limitation.

Le et al. [10]

Method / Dataset Custom CNNs compared with ResNet and EfficientNet transfer learning on Galaxy10 DECaLS.

Advantage EfficientNet achieved about 87%; useful architecture comparison.

Limitation Limited to one dataset and did not address reliability layers.

Fernando et al. [11]

Method / Dataset VGG19 transfer learning on Galaxy10 DECaLS.

Advantage Reported 97.97% accuracy and Macro F1 = 89.68%.

Limitation Accuracy used class-wise one-versus-rest accuracy, so the reported 97.97% is not directly comparable with this framework's sample-level ten-class accuracy. The more balanced Macro F1 comparison slightly favours this framework: 0.8968 versus 0.9073.

❏ **Key Point** — These studies are closest to the hard-label branch because they use Galaxy10 DECaLS and transfer learning; they motivate careful comparison of metrics and protocols.

Self-Supervised and Hybrid Models

Related studies summarised by method, dataset, advantage, and limitation.

Parker et al. [12]

Method / Dataset AstroCLIP self-supervised image–spectrum representation using Galaxy Zoo DECaLS.

Advantage Useful transferable representation evidence.

Limitation Large pretraining cost and weaker performance on complex multi-class questions.

Cao et al. [13]

Method / Dataset CvT hybrid model on Galaxy Zoo / CANDELS-type data.

Advantage Reported around 98% accuracy.

Limitation Different dataset/task; higher complexity and resource needs.

📌 **Key Point** — These studies show the value of representation learning and hybrid attention, but they are not direct Galaxy10 ten-class comparisons.

Galaxy Zoo 2 and Large Survey Catalogues

Related studies summarised by method, dataset, advantage, and limitation.

Urechiatu et al. [14]

Method / Dataset CNN on Galaxy Zoo 2 with preprocessing and augmentation.

Advantage Reported 96.83% accuracy on simpler morphology recognition.

Limitation Struggled with more complicated morphologies.

Bom et al. [15]

Method / Dataset CNN catalog on S-PLUS DR3 for early/late type classification.

Advantage 98.5% purity and large catalog construction.

Limitation Mostly binary early/late task, not detailed ten-class classification.



Key Point — These studies show strong performance in related survey tasks, while also showing that binary or simpler morphology settings are not the same as full ten-class classification.

Attention and Domain Shift

Related studies summarised by method, dataset, advantage, and limitation.

Mao et al. [16]

Method / Dataset DenseNet-SE4 with attention on Galaxy10 DECaLS and GZD-5.

Advantage 88.96% accuracy; designed for visually similar galaxy classes.

Limitation Eight classes, not the original ten-class Galaxy10 setting.

Ye et al. [17]

Method / Dataset Unsupervised domain adaptation from Galaxy Zoo DECaLS to BASS/MzLS.

Advantage Addresses survey shift and target-domain transfer.

Limitation Assumes related underlying distributions.



Key Point — These studies motivate attention-based models and cross-survey caution; they support testing generalisation rather than reporting one dataset score only.

Uncertainty-Aware Limited-Label Learning

Final related study and its direct link to this thesis.

Chen et al. [18]

Method / Dataset Semi-supervised learning with uncertainty estimation on Galaxy10 DECaLS and Galaxy Zoo 2.

Advantage 76.8% with 2,000 labelled samples; uncertainty-aware pseudo-label control.

Limitation Limited-label setting cannot be compared directly with fully supervised results.

Link with this work

- Uncertainty is useful when labels are limited or predictions are ambiguous.
- This thesis uses confidence and error behaviour as supporting reliability evidence.
- Limited-label results are not compared directly with fully supervised held-out tests.



Key Point — The study closes the related-work review by connecting uncertainty-aware learning to our reliability-oriented framework and its careful comparison discipline.

Related Work Summary and Research Gap

The gap after reviewing the previous studies.

Accuracy focus

Many studies report strong performance but focus mainly on score values.

Separated evidence

Uncertainty, robustness, and explainability are often studied as separate directions.

Different protocols

Datasets, labels, tasks, and metrics differ, so results cannot always be directly compared.

Framework need

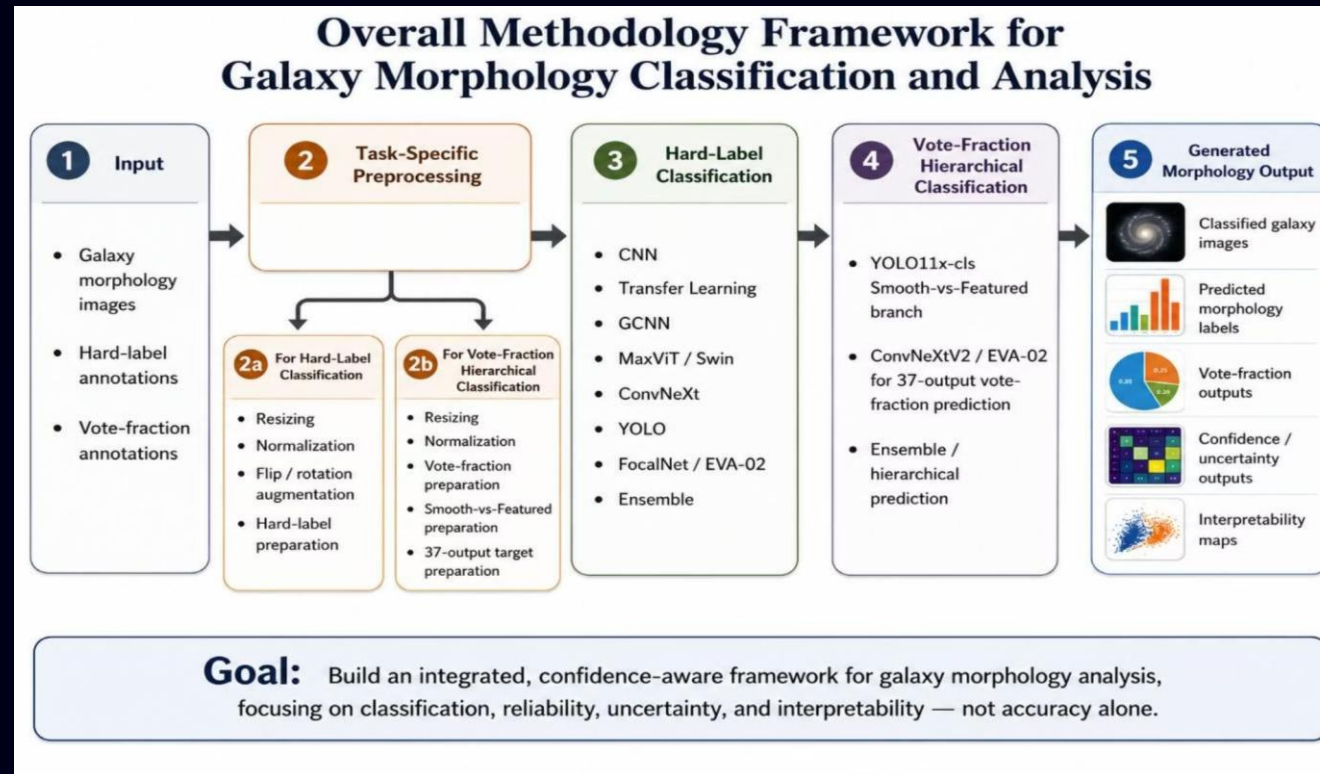
A scientific AI framework should connect prediction, confidence, robustness, and visual explanation.

A framework that connects hard-label classification, confidence-aware routing, vote-fraction prediction, robustness, uncertainty, and visual explanation.

❏ **Key Point** — The gap is not one missing model; it is the absence of one framework that connects performance, confidence, robustness, uncertainty, and explanation while keeping datasets and metrics separated.

Overall Proposed Methodology

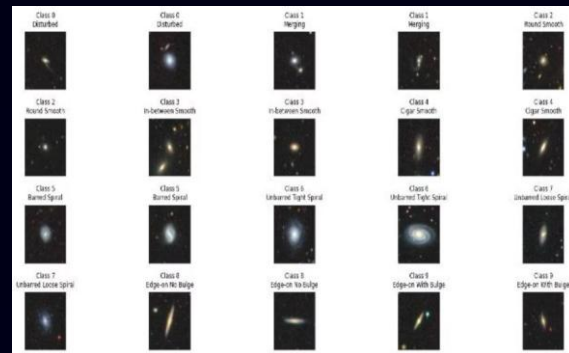
Sequential data flow from input to task-specific modelling and reliability analysis.



Key Point — The framework follows the thesis methodology: data preparation, task-specific modelling, prediction, reliability analysis, and final evaluation.

Input Datasets and Their Roles

Two datasets were selected for complementary hard-label and vote-fraction tasks.



Galaxy10 DECaLS

17,736 labelled images · 10 classes · used for controlled full ten-class hard-label classification and error analysis.

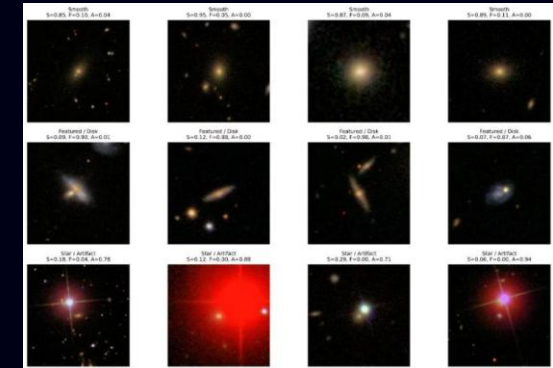
Galaxy10 DECaLS examples for hard-label morphology classification.

Kaggle test role

79,975 unlabelled test images · used only for supporting external scoring after the system was fixed.



Key Point — The datasets are not interchangeable: Galaxy10 supports hard labels, while Galaxy Zoo supports vote fractions and confidence-aware analysis.



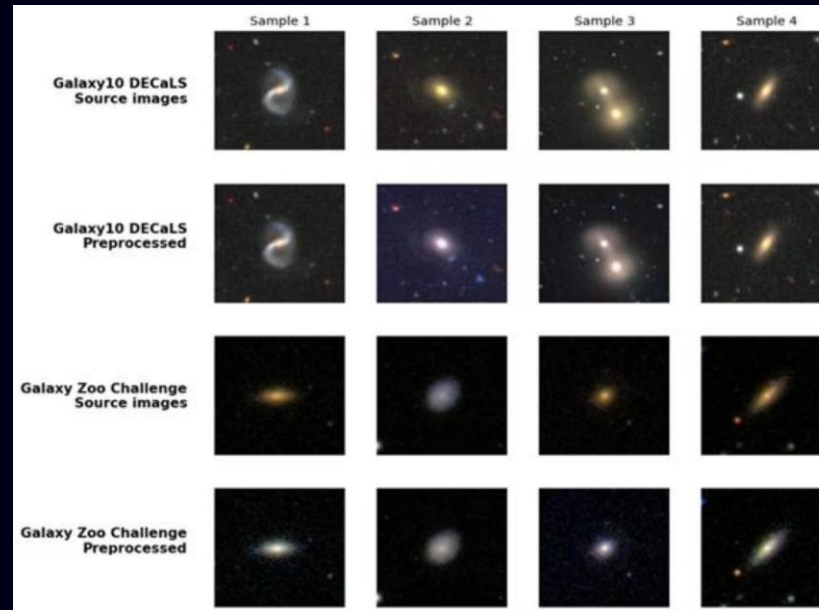
Galaxy Zoo Challenge

61,578 labelled images · 37 vote-fraction targets · used for hierarchical morphology prediction and Smooth-vs-Featured routing.

Galaxy Zoo examples with first-gate vote-fraction information.

Preprocessing by Task

Preprocessing preserved the meaning of each label structure.



Hard-label branch

Images were prepared for ten-class Galaxy10 DECaLS classification while preserving the original class labels.

Vote-fraction branch

Galaxy Zoo labels were treated as 37 continuous outputs rather than converted into one fixed class.

Smooth vs Featured routing

High-agreement first-gate vote fractions were used to construct a binary confidence-aware routing task.

❏ **Key Point** — The same image-processing logic is not blindly applied to each task; preprocessing follows the target structure and evaluation purpose.

Experimental Control and Data Partitioning

How model selection and test evidence were kept separate.

Galaxy10 DECaLS split

Train: 12,415 · Validation: 2,660 · Held-out test: 2,661 · Role: full ten-class classification

Galaxy Zoo split

Train: 43,104 · Validation: 9,237 · Internal held-out test: 9,237 · Role: 37-output prediction

Kaggle evidence

Unlabelled test: 79,975 · Predictions generated after final system freezing · Role: supporting external benchmark

Selection discipline

Validation was used for model, hyperparameter, and training-recipe selection. Held-out test labels remained unseen until the selected system was fixed.

Leakage control

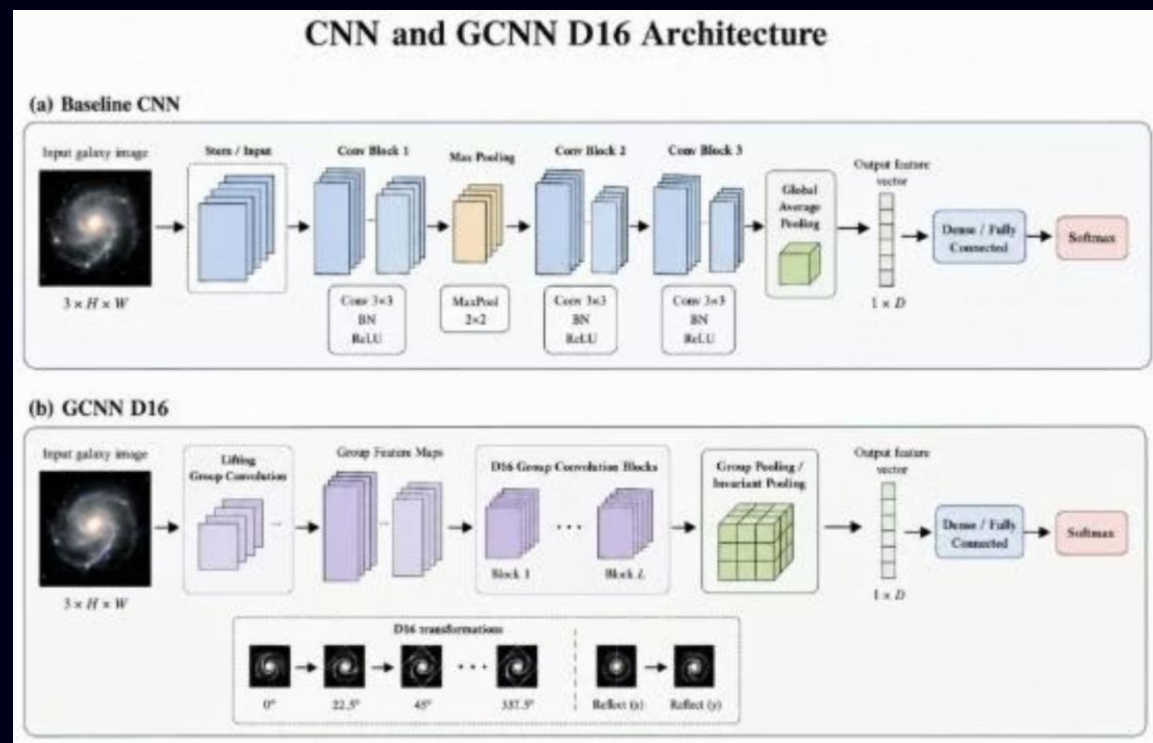
Validation and held-out test images were not used to create augmented training copies or to select final test outcomes.



Key Point — Training, validation, held-out testing, and external Kaggle evidence serve different roles and are not mixed.

CNN and GCNN D16 Family

Baseline and symmetry-aware local morphology feature learning.



Why selected?

CNNs establish a conventional local-feature baseline. GCNN D16 tests whether rotation/reflection equivariance helps because galaxy orientation should not change morphology class.

Role in this research

Baseline and symmetry-aware reference for the hard-label branch; not claimed as the strongest final model.

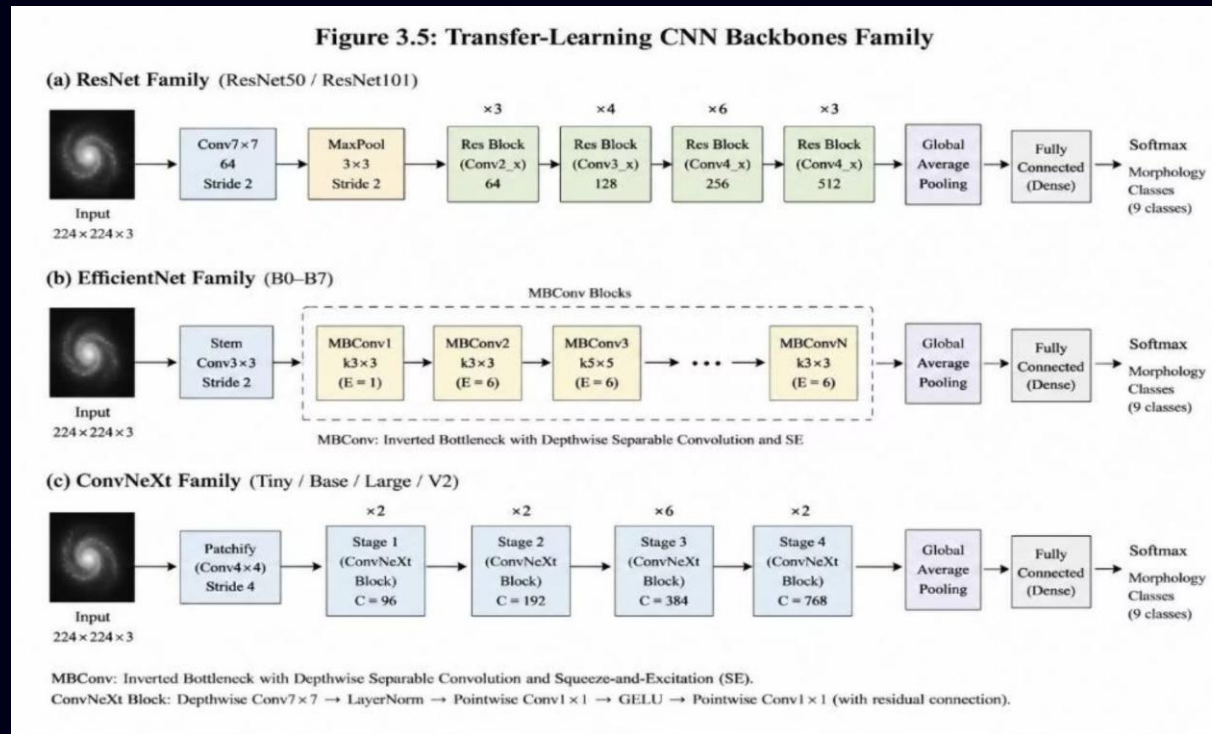
Outcome

GCNN D16 validated the symmetry-aware direction but did not outperform the strongest transfer-learning model.

Key Point — CNN/GCNN was included as a baseline and symmetry-aware reference: orientation-aware modelling was scientifically motivated, but it was not the strongest final model.

Transfer-Learning CNN Backbones

Pretrained visual representations adapted to galaxy images.



Why selected?

Galaxy10 is labelled but not huge compared with general image datasets; pretrained backbones can improve generalisation.

Role in this research

Main candidate family for the full ten-class task and the Galaxy Zoo 37-output branch.

Outcome

ConvNeXtV2 produced the strongest official Galaxy10 held-out result; stronger augmentation did not exceed the cleaner recipe in validation.

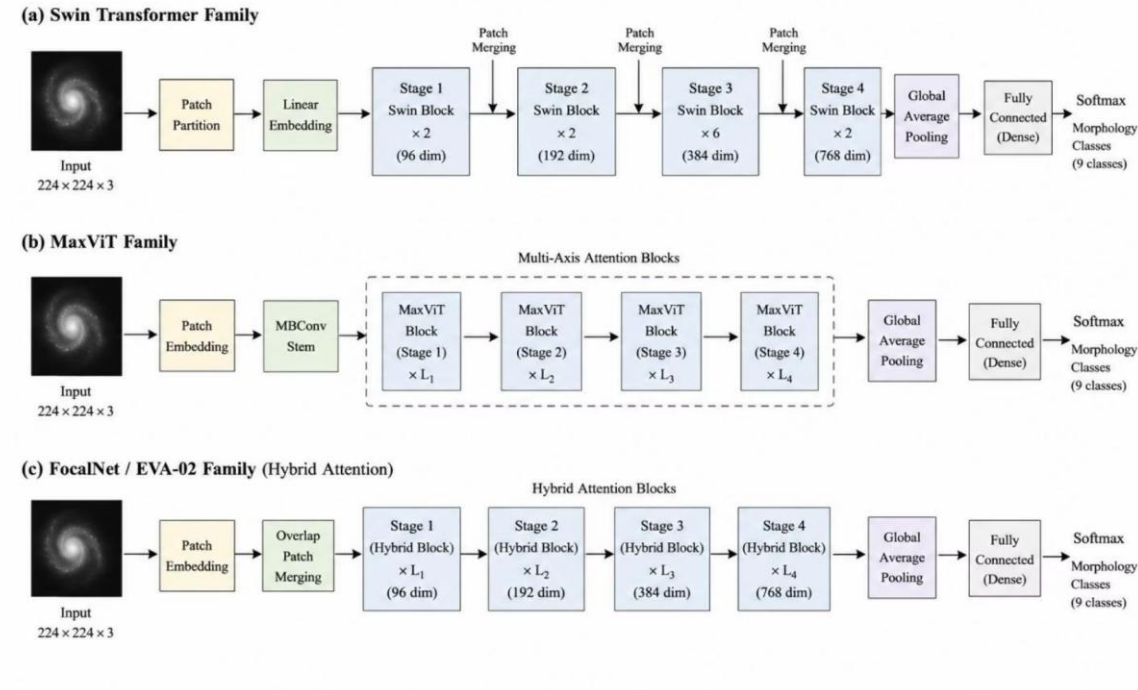


Key Point — Transfer-learning CNNs were the main candidate family; ConvNeXtV2 produced the strongest official full ten-class held-out result.

Transformer and Hybrid Attention Family

Global spatial reasoning for complex morphology.

Figure 3.6: Transformer and Hybrid Attention Family



Why selected?

Galaxy morphology may depend on relations between the core, disk, arms, and broader structure, not only local features.

Role in this research

Tested whether attention and hybrid representations improve separation of visually related classes.

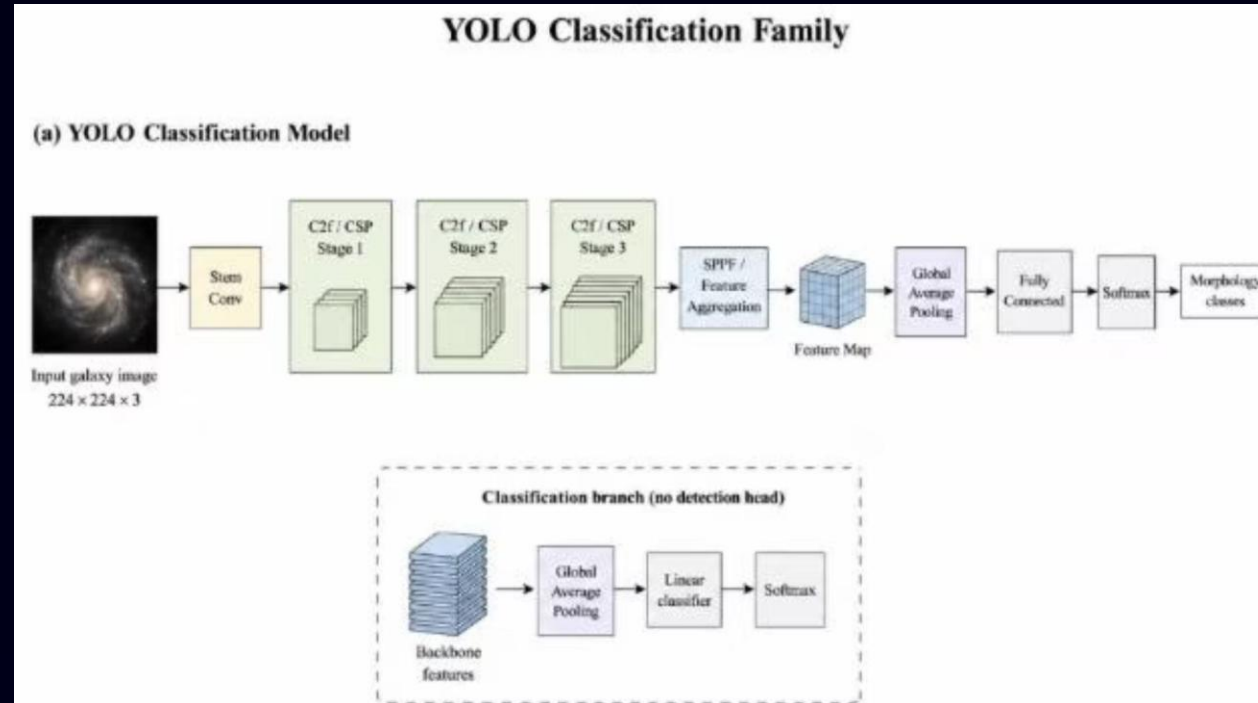
Outcome

Swin Large and MaxViT provided useful validation evidence; MaxViT improved with TTA, but size alone did not guarantee the best result.

□ **Key Point** — Transformer and hybrid models tested broader spatial reasoning, but larger attention-based models did not automatically replace the selected ConvNeXtV2 result.

YOLO Classification Family

Efficient image-level classifiers for hard-label and routing tasks.



Why selected?

YOLO was used in classification mode, not detection mode; no bounding boxes were required.

Role in this research

Independent image-level classifier family for Galaxy10 and high-confidence Smooth-versus-Featured routing.

Outcome

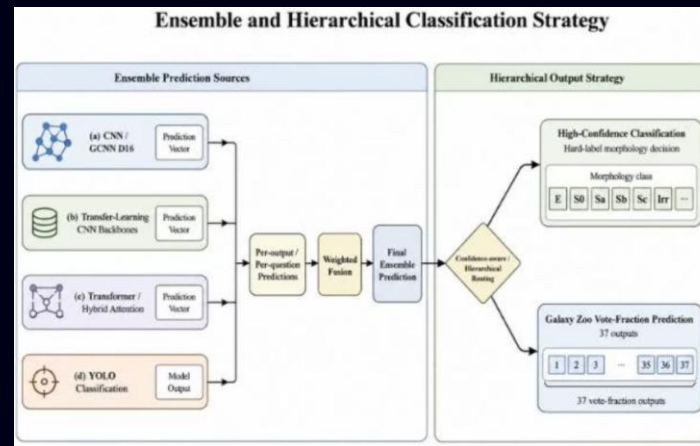
YOLO benefited from RandAugment/TTA/probability ensemble in validation and achieved the strongest Smooth-versus-Featured held-out routing result.



Key Point – YOLO was used in classification mode, not detection mode; it provided independent hard-label evidence and the strongest high-confidence routing result.

Ensemble and Hierarchical Strategy

Connecting model families with task-aware decision flow.



Why selected?

Different model families can make different errors, so fusion tests whether complementary representations improve prediction.

Role in this research

Weighted fusion, specialists, and hierarchical routing connect hard labels, high-confidence decisions, and 37-output vote fractions.

Outcome

Reduced seven-class fusion is diagnostic evidence only; it is not a replacement for the official full ten-class held-out test.

Key Point — Ensemble and hierarchical experiments tested complementary behaviour; reduced seven-class evidence remained diagnostic and did not replace the full ten-class evaluation.

Training and Evaluation Design

Task-specific losses, disciplined model selection, and reliability evidence.

Hard-label classification

Cross-entropy loss · one class per image · accuracy, precision, recall, Macro F1, confusion matrix.

Smooth vs Featured

Binary image classification using high-agreement first-gate labels · held-out accuracy and Macro F1.

37-output prediction

Continuous multi-output prediction · RMSE, MAE, per-output analysis, first-gate interpretation.

Training and generalisation strategies

Training-time augmentation, regularisation, optimisation/training control, and inference-time generalisation were evaluated separately: AdamW, scheduling, checkpoint selection, TTA, and probability averaging.

Reliability analysis

Controlled perturbations, confidence and error behaviour, Grad-CAM, and saliency maps were used as supporting reliability evidence.

Reporting discipline

Every result is labelled as validation, held-out test, reduced-class diagnostic, or external Kaggle evidence.



Key Point — Metrics were selected by output type, while augmentation and generalisation effects were interpreted by model and task rather than treated as one universal improvement.

Hard-Label Results Across Model Families

Representative evidence from each family, with evaluation status kept explicit.

Official held-out result

ConvNeXtV2 · Full 10-class task · Test Acc = 0.9158 · Macro F1 = 0.9073 · **Selected final model**

Validation evidence

GCNN D16: Acc 0.8865, F1 0.8782 · Swin Large 384: Acc 0.8850, F1 0.8628 · YOLO11 ensemble: Top-1 0.8932, F1 0.8787 · Used for family comparison

Reduced-class diagnostic

YOLO + FocalNet · Reduced 7-class · Val Acc = 0.9803 · F1 = 0.9709 · **Diagnostic only — not full 10-class**

 **Key Point** — The official held-out ConvNeXtV2 result, validation evidence, and reduced-class diagnostic result are separated to avoid reading the slide as one leaderboard.

Detailed ConvNeXtV2 Ten-Class Evaluation

Held-out performance, error concentration, and representative predictions.

91.58%

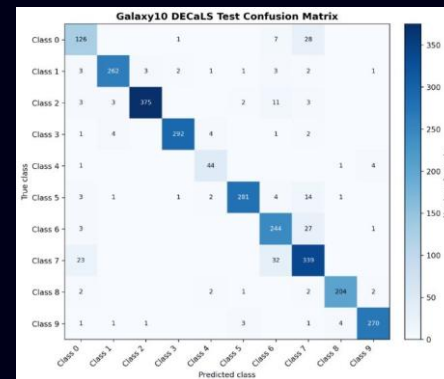
Held-out test accuracy

0.9073

Macro F1

2,661

Held-out test images



Galaxy10 DECaLS ten-class confusion matrix.



Representative correct and incorrect ConvNeXtV2 predictions.

Key error pattern

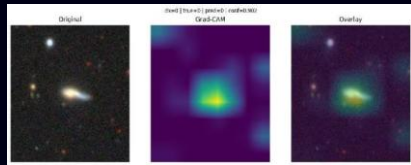
Main confusions: Class 0 (Disturbed), Class 6 (Unbarred Tight Spiral), and Class 7 (Unbarred Loose Spiral). This is hard-label ambiguity, not uniform failure.



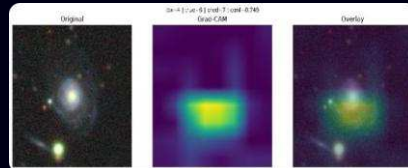
Key Point — The confusion matrix explains the score: most classes are strong, while Disturbed and the two unbarred spiral classes account for the main hard-label ambiguity.

Explainability, Robustness, and Uncertainty for Galaxy10

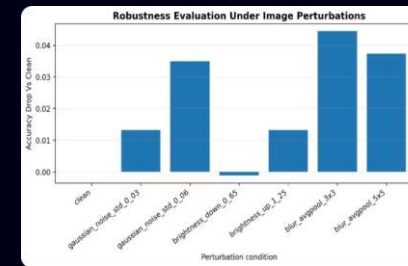
Three checks beyond the held-out accuracy.



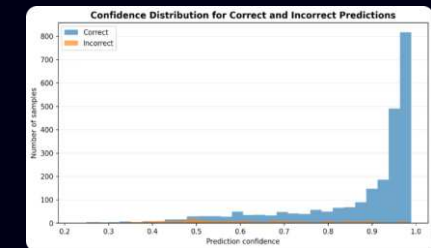
Grad-CAM example: lower-error prediction.



Grad-CAM example: higher-error prediction.



Robustness under controlled perturbations.



Confidence distributions for correct and incorrect predictions.

Interpretation

Noise and blur produced the clearest degradation, and incorrect cases generally showed weaker confidence. Grad-CAM was used as supporting visual inspection evidence.

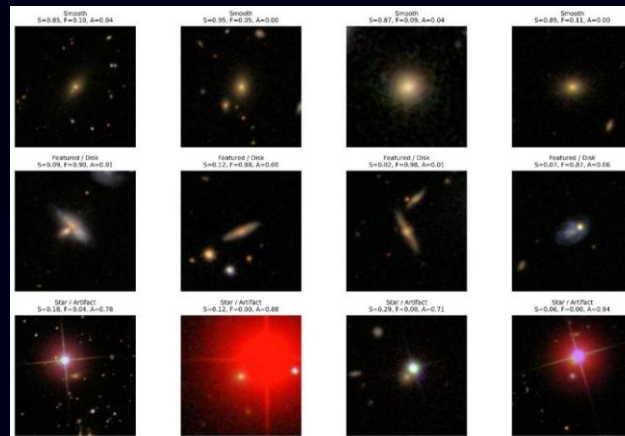
Key Point — Reliability evidence identifies when the model is likely to be less dependable; it does not claim that saliency maps prove causal understanding.

High-Confidence Smooth vs Featured Classification

A first-gate routing task built from high-agreement Galaxy Zoo vote fractions.

99.17%

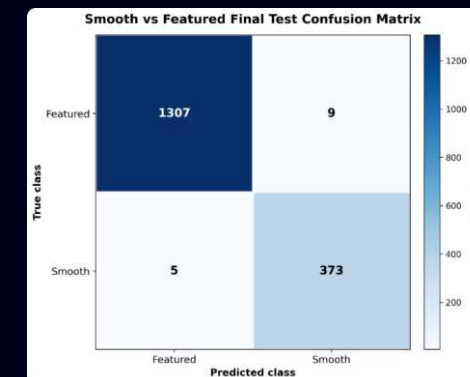
Held-out accuracy



High-agreement examples used to construct Smooth and Featured labels.

0.9881

Macro F1



Official held-out confusion matrix.

Interpretation

The result demonstrates reliable routing on a high-confidence binary subset. It must not be compared directly with the full ten-class Galaxy10 task.

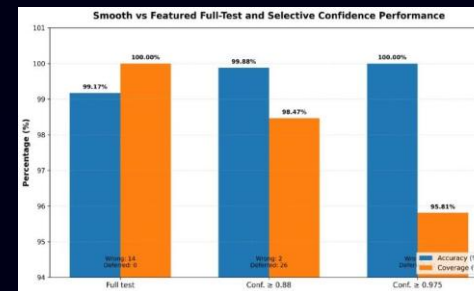
Key Point — The 99.17% result applies only to high-agreement Smooth-versus-Featured routing—not to the full ten-class morphology problem.

Smooth vs Featured: Errors, Confidence, and Robustness

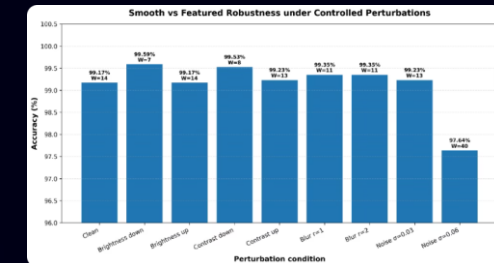
The binary result was evaluated beyond one accuracy number.



Representative correct and incorrect predictions.



Full-test and selective-confidence performance.



Robustness under controlled perturbations.

What the additional analysis shows

High confidence increases selective accuracy, but stronger perturbations still reduce performance. The classifier is strong on the selected label definition, not invulnerable.

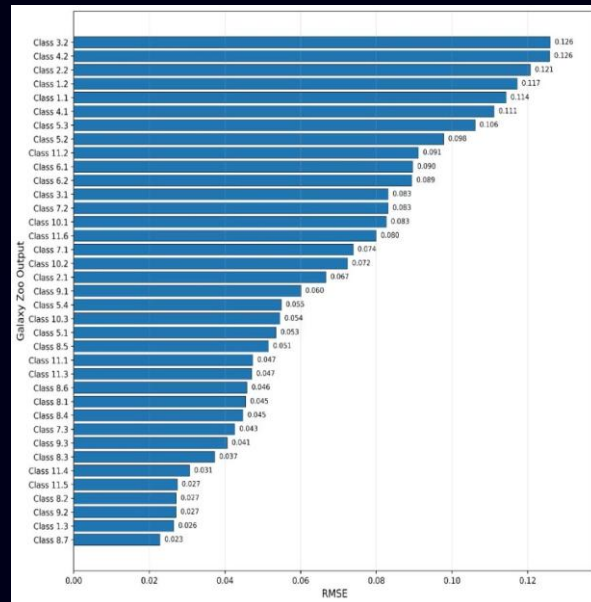
Key Point — Selective confidence is useful as a routing rule because it allows uncertain cases to be deferred rather than forced into a high-confidence decision.

Galaxy Zoo 37-Output Vote-Fraction Prediction

Continuous morphology prediction on the internal held-out test.

0.074912

Held-out RMSE



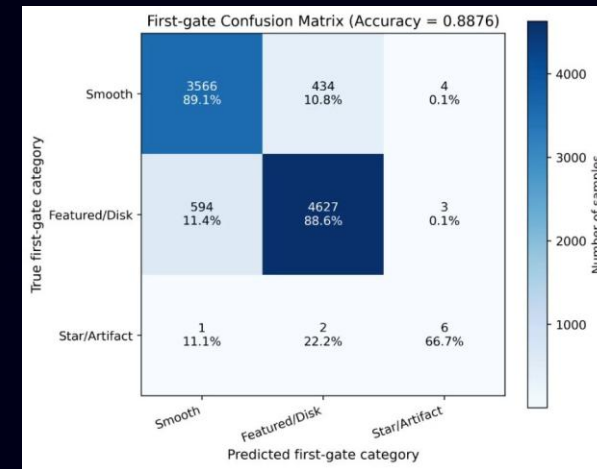
Per-output RMSE across the 37 vote-fraction targets.

0.042315

Held-out MAE

0.8876

First-gate dominant-class agreement



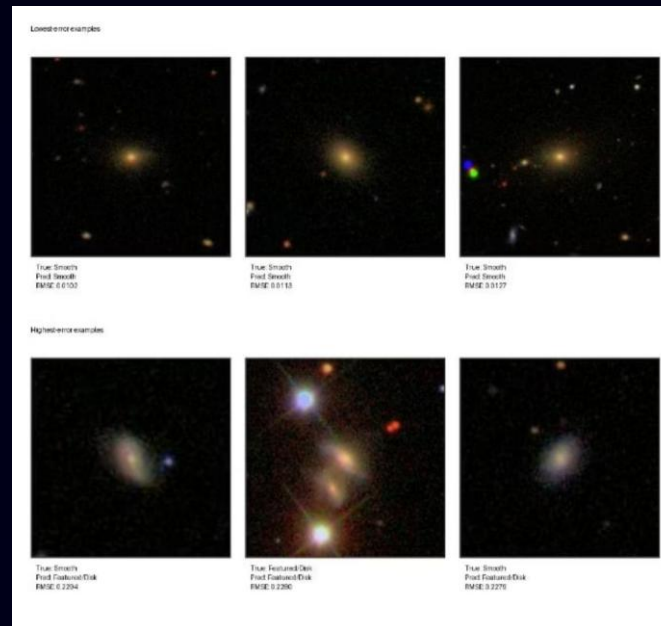
First-gate prediction analysis for the 37-output model.



Key Point — RMSE and MAE are used because the outputs are continuous vote fractions. Seven-view TTA improved validation RMSE from 0.07608 to 0.07469, and lower RMSE means better prediction.

Qualitative Error Analysis for Vote Fractions

Representative lowest-error and highest-error predictions.



Lowest-error and highest-error qualitative prediction panels.

Lowest-error cases

Predicted vote fractions closely matched the target hierarchy, usually for visually clearer morphologies.

Highest-error cases

Larger errors were associated with ambiguous, mixed, or difficult visual structures and disagreement across outputs.

Why this matters

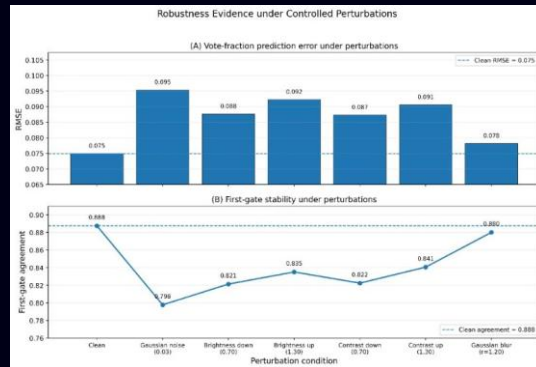
Sample-level examples connect aggregate RMSE to actual prediction behaviour and prevent the metric from being interpreted without context.



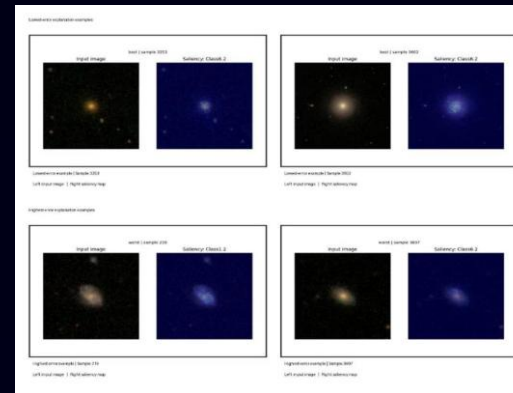
Key Point — The model is evaluated at both aggregate and sample levels; qualitative examples show what a low or high vote-fraction error means in practice.

Framework-Level Reliability Evaluation

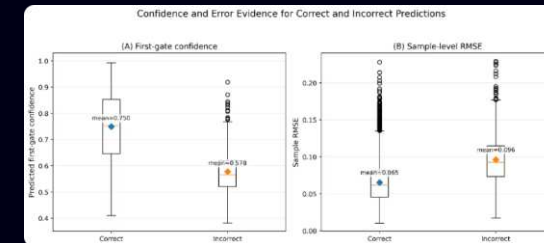
Robustness, confidence, error behaviour, and visual explanation considered together.



Framework-level robustness evidence.



Saliency evidence for selected Galaxy Zoo cases.



Confidence and error behaviour.

Robustness finding

Stable under moderate perturbations; stronger perturbations reduced performance.

Confidence / error finding

Lower-confidence and higher-error cases can be identified for careful review.

Visual explanation finding

Grad-CAM and saliency support visual inspection, but do not guarantee correctness.

❏ **Key Point** — Framework-level reliability is supported by combined evidence from robustness, confidence/error behaviour, and visual explanation; no single score guarantees correctness.

Discussion and Comparison with Previous Studies

Interpreting results by task, metric definition, and evaluation protocol.

Galaxy10 official result

ConvNeXtV2 held-out test: 91.58% sample-level ten-class accuracy and Macro F1 = 0.9073.

Fernando et al. [11]

Reported 97.97% by averaging class-wise one-versus-rest accuracies; Macro F1 = 0.8968.

High-Confidence Smooth vs Featured

YOLO11x achieved 99.1736% held-out accuracy and Macro F1 = 0.9881 on the high-confidence binary morphology-routing task.

Dieleman et al. [2]

A rotation-invariant CNN with model averaging achieved a Private RMSE of 0.07492 for Galaxy Zoo 37-output vote-fraction prediction and more than 99% accuracy for most high-agreement questions.

Galaxy Zoo benchmark evidence

Held-out RMSE = 0.074912; supporting Kaggle Private RMSE = 0.07440 versus original benchmark 0.07492.

Fernando et al. used class-wise one-versus-rest accuracy, so the reported 97.97% is not directly comparable with this framework's sample-level ten-class accuracy. The more balanced Macro F1 comparison slightly favours this framework: 0.8968 versus 0.9073.

 **Key Point** — The framework delivered strong task-specific results: 91.58% for full ten-class classification, 99.1736% for high-confidence routing, and a Kaggle Private RMSE of 0.07440 versus the 0.07492 Galaxy Zoo benchmark.

Conclusion, Limitations, and Future Work

Final synthesis of the thesis.

Conclusion

- Deep learning can support galaxy morphology classification.
- Predictive performance must be complemented with confidence, robustness, vote-fraction modelling, and visual explanation evidence.

Main Contributions

- Structured model-family evaluation
- Confidence-aware extension
- Robustness and uncertainty evidence
- Grad-CAM and saliency inspection

Limitations

- Findings are tied to the selected datasets and protocols.
- Robustness tests were controlled, not exhaustive.
- Visual explanations were mainly qualitative.
- The framework supports interpretation; it does not replace it.

Future Work

- Cross-survey validation
- Confidence calibration
- Probabilistic and Bayesian uncertainty
- Realistic survey effects
- Human-in-the-loop review for ambiguous cases

Thank You

Questions and Discussion



Key Point — The thesis contribution is a disciplined confidence-aware AI evaluation framework for scientific image analysis, not a claim that one model is universally best.