



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

Study Summary

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

Problem

The increasing number of imaged galaxies makes manual morphology classification impractical and time-consuming 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

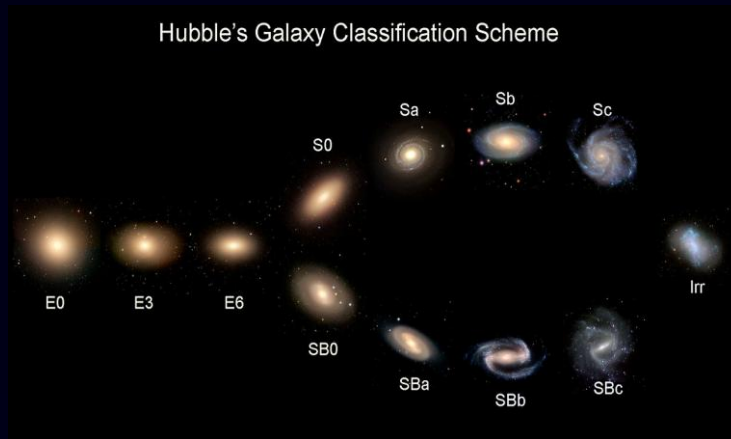
Accuracy, precision, recall, Macro F1, RMSE, MAE, controlled perturbations, confidence analysis, Grad-CAM, and saliency maps.



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

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.



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.

Central Challenge

Accurate prediction alone does not provide sufficient evidence of robustness, confidence, ambiguity awareness, or visual interpretability for scientific use.



Key Point — The central gap is the absence of a unified framework that connects predictive performance with robustness, confidence, uncertainty-related evidence, and visual explanation.

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.	Extend the framework from hard-label classification to high-confidence routing and continuous vote-fraction prediction.
04	05	06
Evaluate classification using accuracy and Macro F1, and continuous multi-output prediction using RMSE and MAE.	Evaluate robustness under controlled image perturbations and analyse confidence and error behaviour.	Use Grad-CAM and saliency maps to inspect whether decisions focus on relevant structures.

 **Key Point** — Each objective is linked to a corresponding methodological component and evaluation presented 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.

Related Work Overview and Research Gap

The gap after reviewing the 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, uncertainty-related evidence, and visual explanation.

Identified Gap

The identified gap is the absence of an integrated framework connecting hard-label classification, confidence-aware routing, vote-fraction prediction, robustness, uncertainty-related evidence, and visual explanation.

- ❏ **Key Point** — The gap is not one missing model; it is the absence of an integrated framework that connects performance, confidence, robustness, uncertainty-related evidence, and explanation while preserving clear distinctions between datasets, tasks, and metrics.

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 Performance remains dependent on the quality, consistency, and representativeness of volunteer-provided annotations.

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 directly motivate two components of our framework: continuous 37-output vote-fraction prediction for Galaxy Zoo and the evaluation of models designed to reduce sensitivity to galaxy orientation.

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 post-hoc visual explanation evidence but does not establish causal model reasoning or a new performance 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 The reported result was obtained on a different dataset and protocol, without an integrated robustness, confidence, and explainability framework.



Key Point — Bhambra et al. motivate our use of Grad-CAM and saliency maps, while Domínguez Sánchez et al. support the CNN baseline; together, they show why our framework evaluates both predictive performance and visual explanation rather than accuracy alone.

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 semi-supervised GAN-based training was reported to be unstable, which may affect consistency and reproducibility.

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 Galaxy10 studies motivate our controlled evaluation across multiple model families and classes: label-efficient learning and augmentation may help, but their benefits must be validated rather than assumed to generalise uniformly.

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 Reported about 87% accuracy with EfficientNet and provided a useful comparison between custom and pretrained CNN architectures.

Limitation The study was limited to one dataset and did not integrate robustness, confidence, uncertainty-related evidence, or visual explanation.

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 Galaxy10 studies directly motivate our hard-label branch and evaluation of pretrained CNN backbones; they also demonstrate why results must be interpreted using matched task definitions and metrics, with Macro F1 providing the fairer comparison to our ten-class model.

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 Provides evidence that self-supervised image–spectrum representations can support transfer to downstream galaxy-analysis tasks.

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 — Parker et al. motivate transferable representation learning, while Cao et al. support our evaluation of hybrid attention models; in our framework, these ideas were examined through pretrained and transformer/hybrid model families under a consistent Galaxy10 protocol rather than through direct comparison of reported percentages.

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 — Urechiatu et al. support our use of preprocessing and controlled augmentation, while Bom et al. demonstrate the scalability of CNN-based morphology catalogues; their simpler or binary tasks also motivate our separation of the high-confidence Smooth-versus-Featured branch from the more difficult Galaxy10 ten-class task.

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 — Mao et al. motivate our evaluation of transformer and hybrid attention models for visually similar Galaxy10 classes, while Ye et al. highlight that robustness within one dataset is not equivalent to cross-survey generalisation; our framework tests controlled perturbations but does not claim full domain adaptation.

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

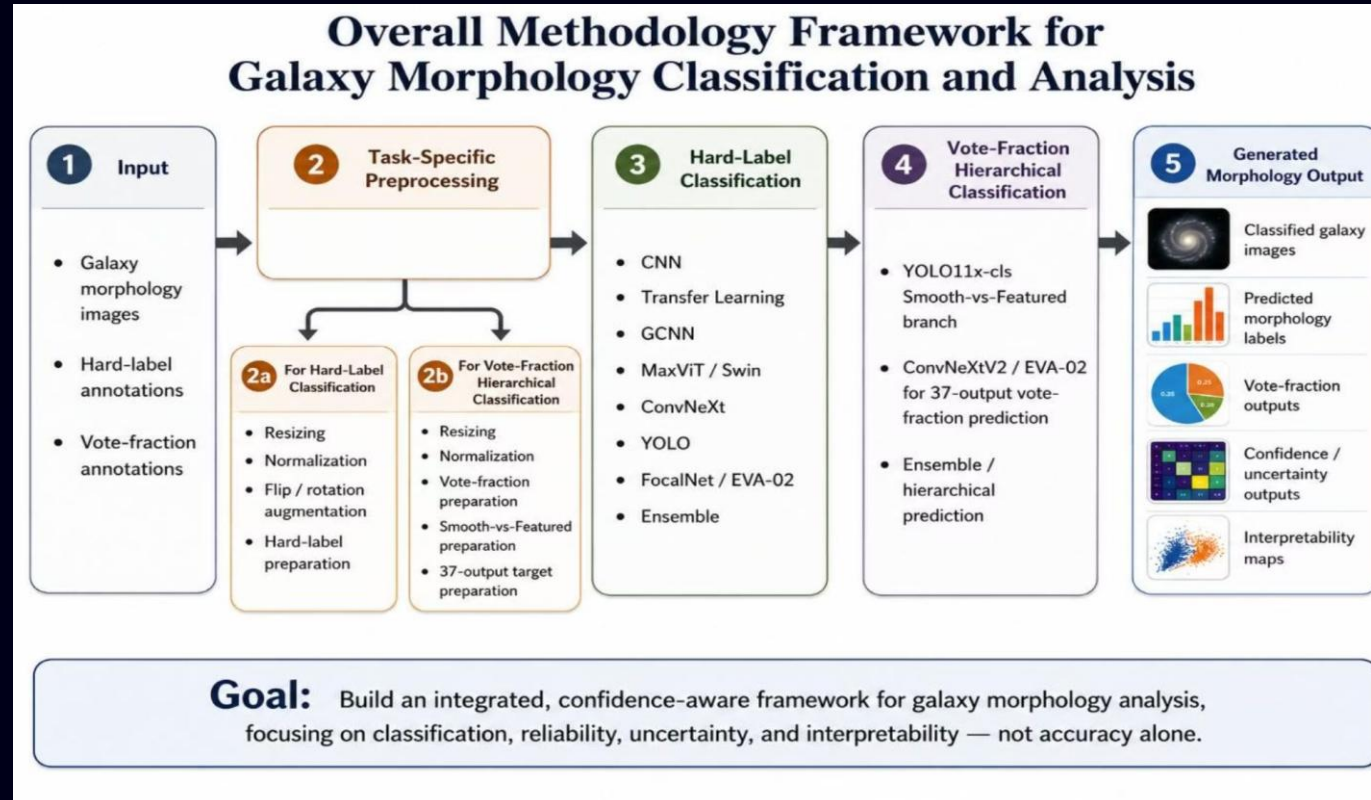
- Uncertainty is useful when labels are limited or predictions are ambiguous.
- This thesis uses confidence and error behaviour as supporting reliability evidence.
- Limited-label results are not compared directly with fully supervised held-out tests.



Key Point — The study directly motivates our use of confidence and error behaviour to identify ambiguous or lower-reliability predictions, while reinforcing that limited-label semi-supervised results must remain separate from our fully supervised held-out evaluations.

Overall Proposed Methodology

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



Key Point — The framework connects three separately evaluated tasks—ten-class hard-label classification, high-confidence Smooth-versus-Featured routing, and 37-output vote-fraction prediction—with task-specific metrics and reliability analyses.

Input Datasets and Their Roles

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



Galaxy10 DECaLS

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

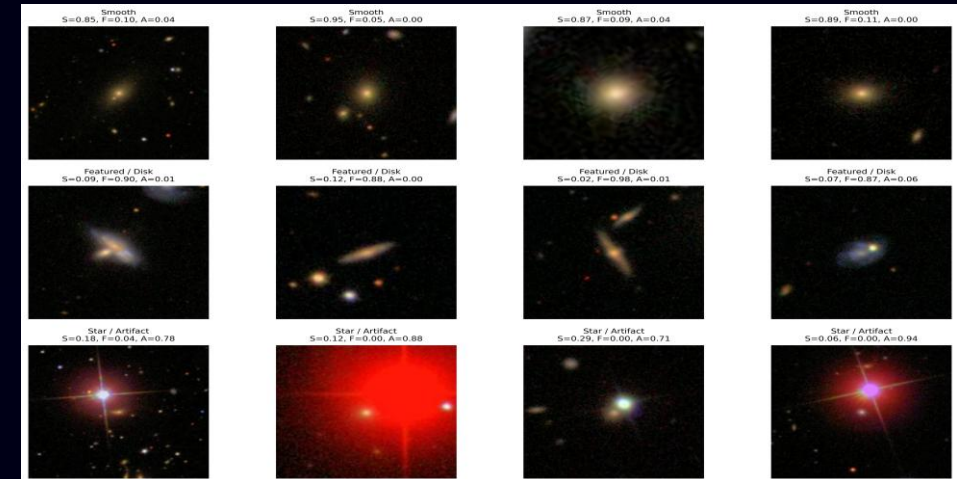
Galaxy10 DECaLS examples for hard-label morphology classification.

Kaggle test role

79,975 unlabelled test images, used only for supporting external scoring after model selection and system configuration were fixed.



Key Point — Galaxy10 provides the controlled ten-class hard-label benchmark, Galaxy Zoo provides high-confidence routing and 37-output continuous targets, and the Kaggle test set provides external scoring only after model development was complete.



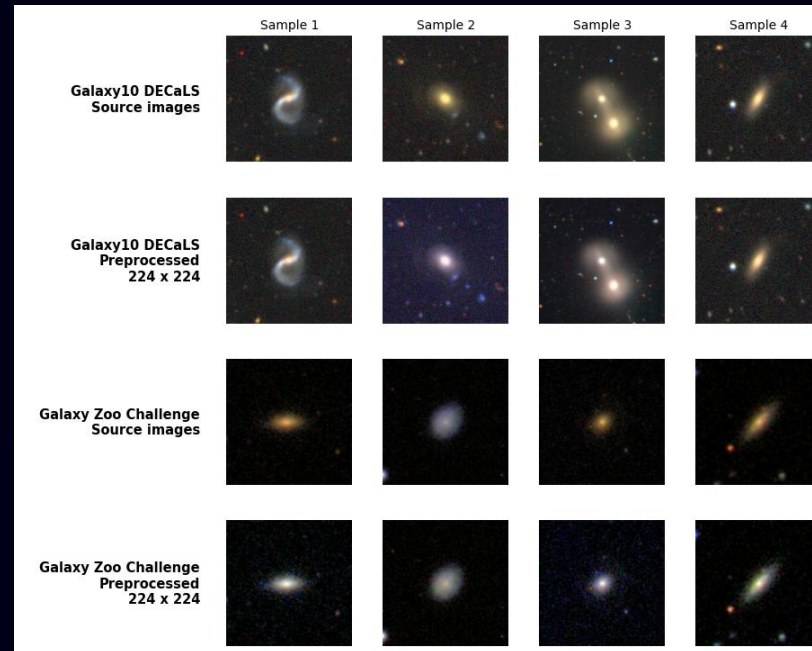
Galaxy Zoo Challenge

61,578 labelled images with 37 vote-fraction targets, used for high-confidence Smooth-versus-Featured routing and continuous multi-output vote-fraction prediction.

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

Task-Specific Preprocessing and Target Preparation

Preprocessing preserved the meaning of each label structure.



Hard-label branch

Galaxy10 DECaLS images were prepared for ten-class classification while preserving their original hard-label class assignments.

Vote-fraction branch

Galaxy Zoo annotations were preserved as 37 continuous vote-fraction targets rather than collapsed into one fixed hard class.

Smooth vs Featured routing

High-agreement first-gate vote fractions were used to construct the binary Smooth-versus-Featured routing task.

❑ **Key Point** — Our framework preserves each task's target structure: Galaxy10 remains a ten-class hard-label problem, Galaxy Zoo remains a 37-output continuous prediction problem, and high-agreement first-gate labels form the Smooth-versus-Featured routing branch.

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 · Used only after model selection and system configuration were fixed · 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

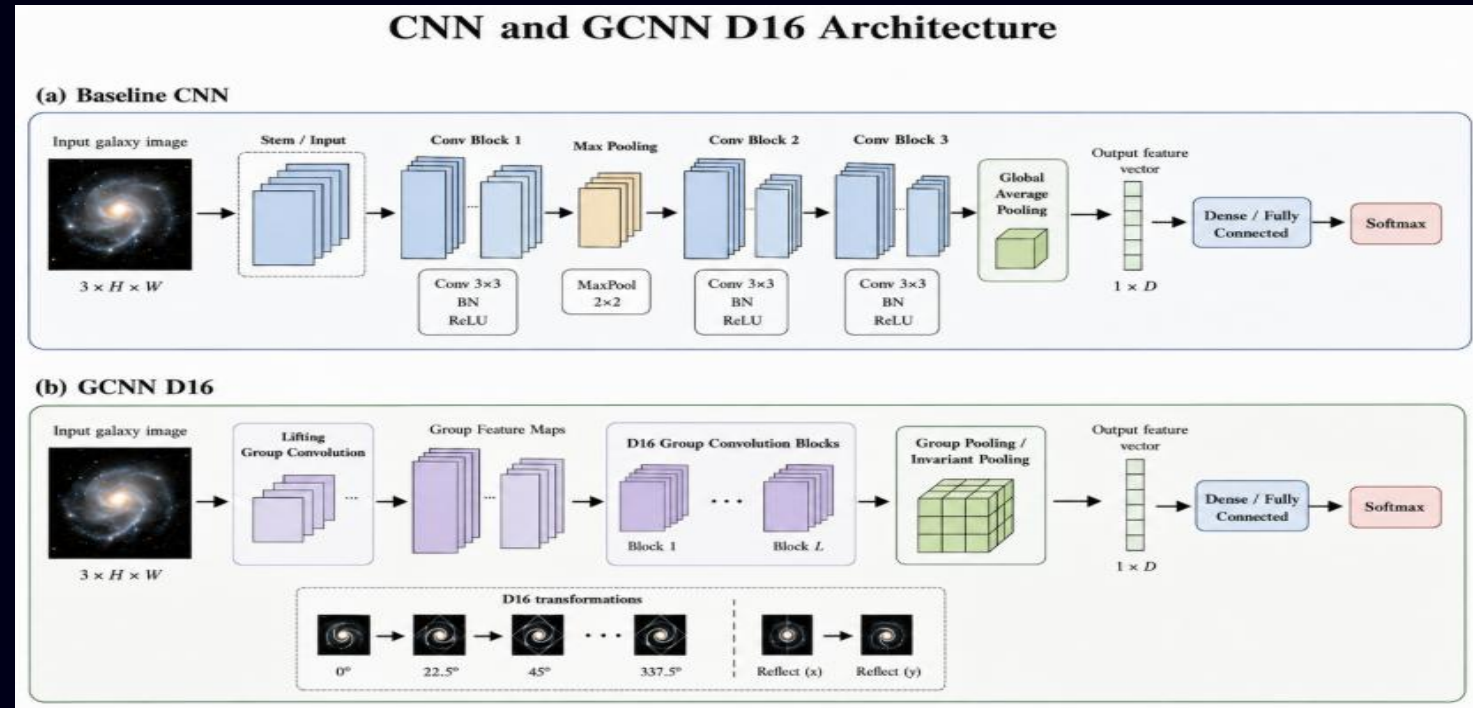
Validation and held-out test images were excluded from training augmentation and held-out test results were not used for model selection.



Key Point — Training, validation, internal held-out testing, and external Kaggle scoring were kept strictly separate to prevent leakage and preserve unbiased final evaluation.

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.

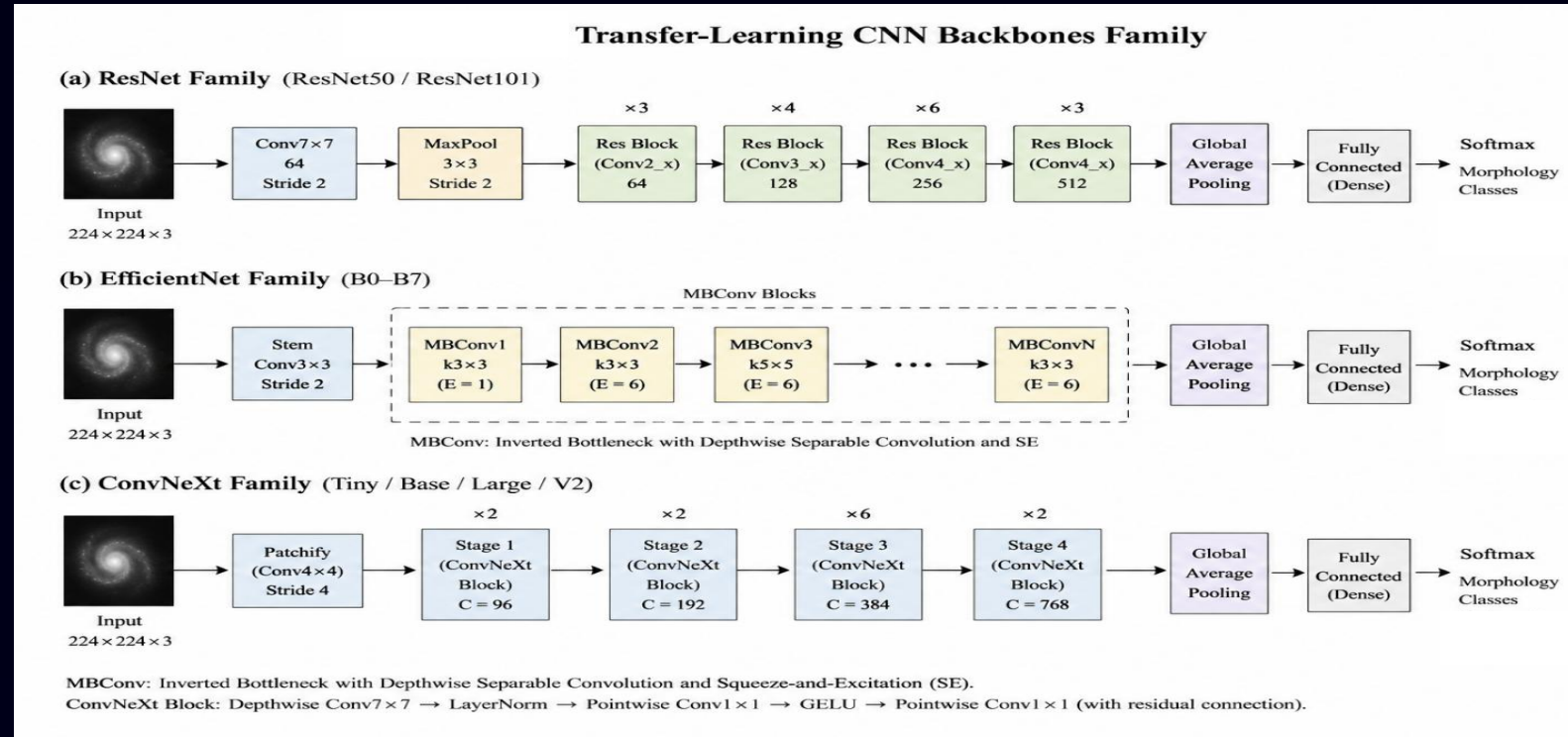
Outcome

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

Key Point — CNN and GCNN D16 established the conventional and symmetry-aware baselines for the hard-label branch; rotation/reflection equivariance was scientifically justified, but the strongest final held-out result came from transfer learning.

Transfer-Learning CNN Backbones

Pretrained visual representations adapted to galaxy images.



Why selected?

Galaxy10 is a moderate-sized labelled dataset; pretrained backbones can transfer general visual representations and improve generalisation.

Role in this research

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

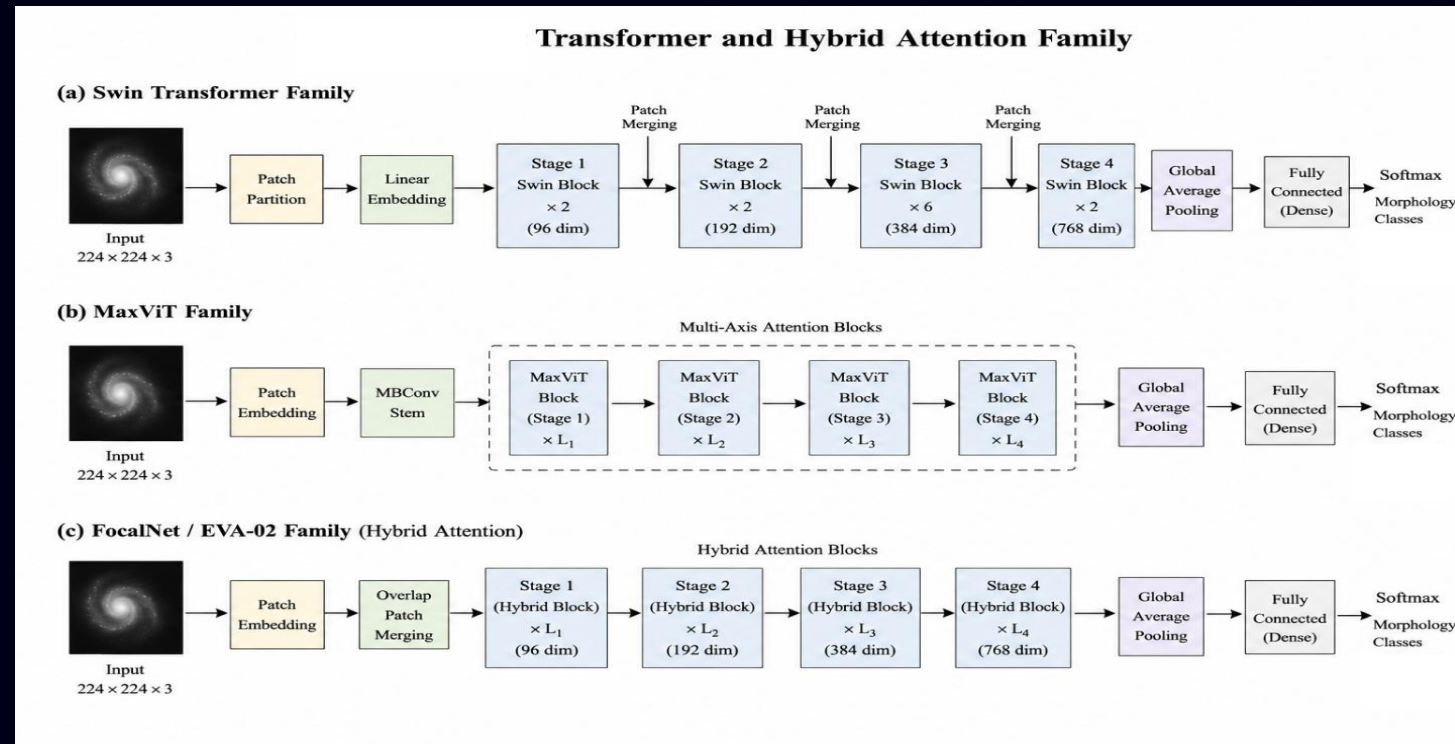
Outcome

ConvNeXtV2 produced the strongest official Galaxy10 held-out result; more aggressive augmentation did not outperform the selected simpler training recipe on validation.

□ **Key Point** – Transfer-learning CNNs were the main candidate family, and ConvNeXtV2 produced the strongest official full ten-class held-out result under the controlled evaluation protocol.

Transformer and Hybrid Attention Family

Global spatial reasoning for complex morphology.



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.

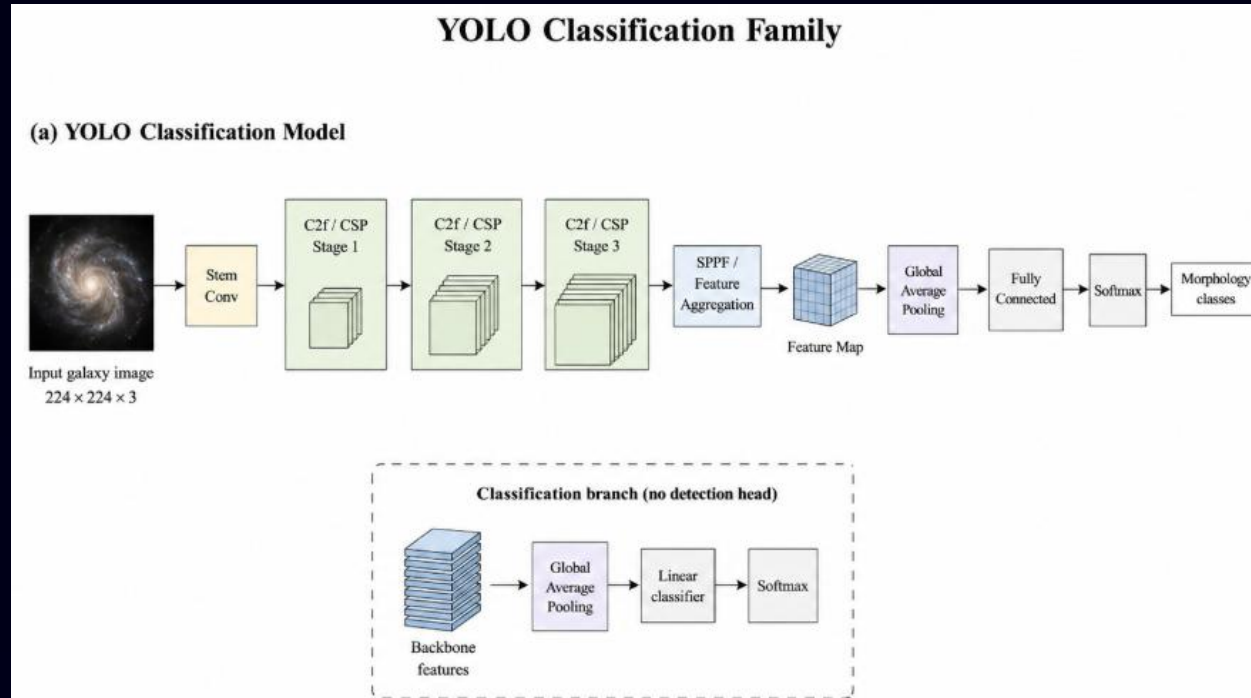
Outcome

Swin Large and MaxViT provided strong validation evidence; MaxViT improved with TTA, but increased attention-model size did not guarantee the strongest 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.

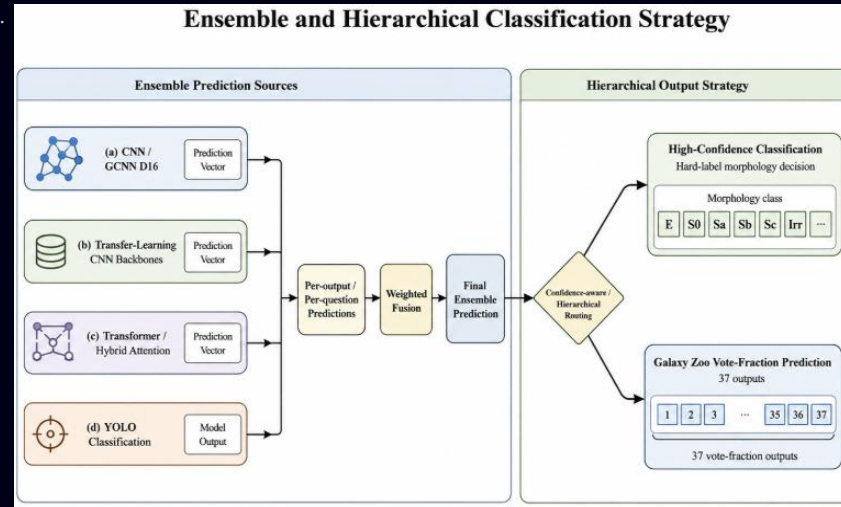
Outcome

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 evaluation, 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, AdamW optimisation, learning-rate scheduling, and checkpoint selection were evaluated separately from inference-time 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

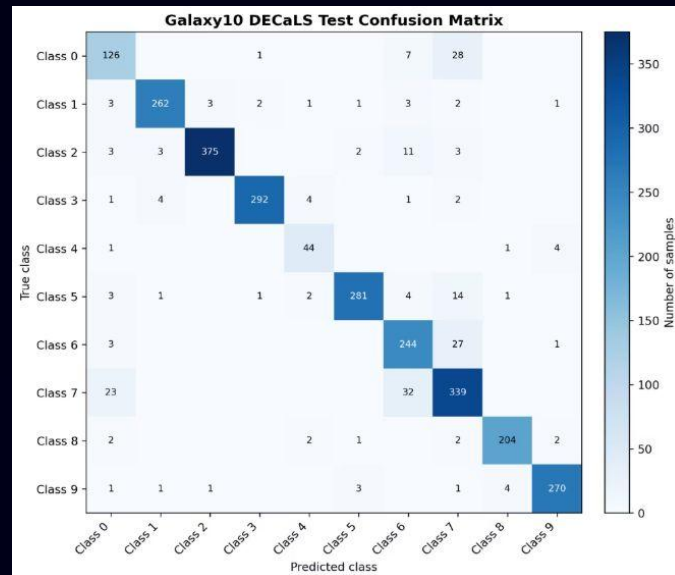
YOLO + FocalNet · Reduced 7-class · Val Acc = 0.9803 · F1 = 0.9709 · **Diagnostic only — not full 10-class**



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.



Galaxy10 DECaLS ten-class confusion matrix.

91.58%

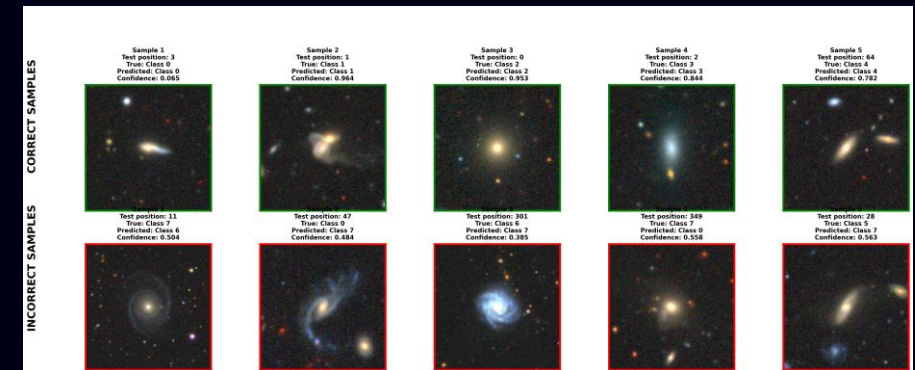
Held-out test accuracy

0.9073

Macro F1

2,661

Held-out test images



Representative correct and incorrect ConvNeXtV2 predictions.

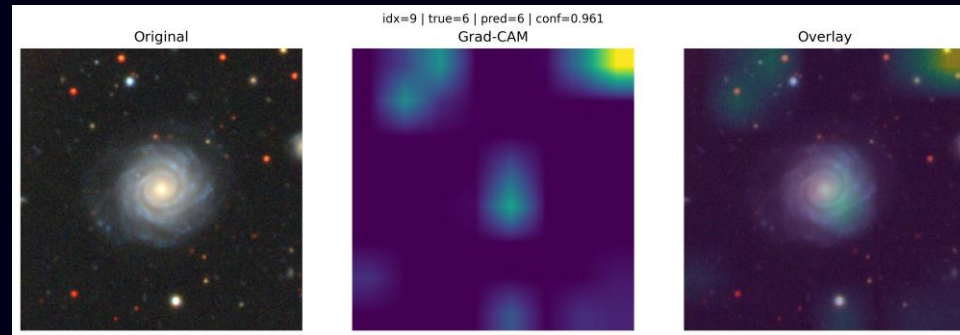
Key error pattern

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

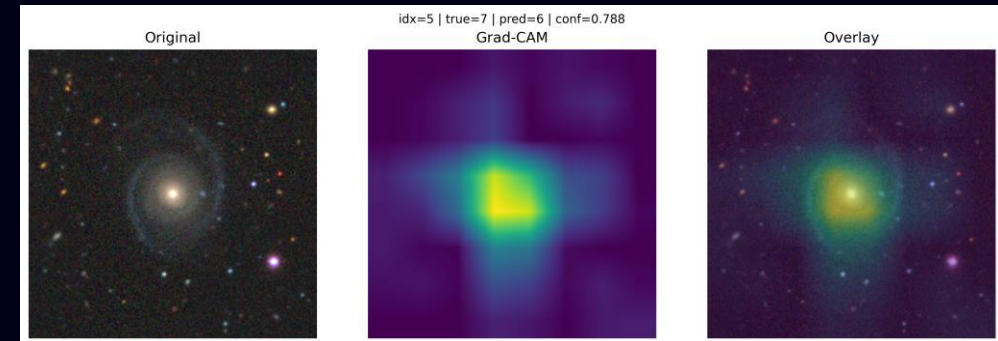
Key Point — The confusion matrix explains the score: most classes are strong, while Disturbed and the two unbarred spiral classes account for the main hard-label ambiguity.

Explainability for Galaxy10: Grad-CAM Analysis

Visual inspection of image regions associated with correct and incorrect predictions.



Grad-CAM example: correct prediction.



Grad-CAM example: misclassified prediction.

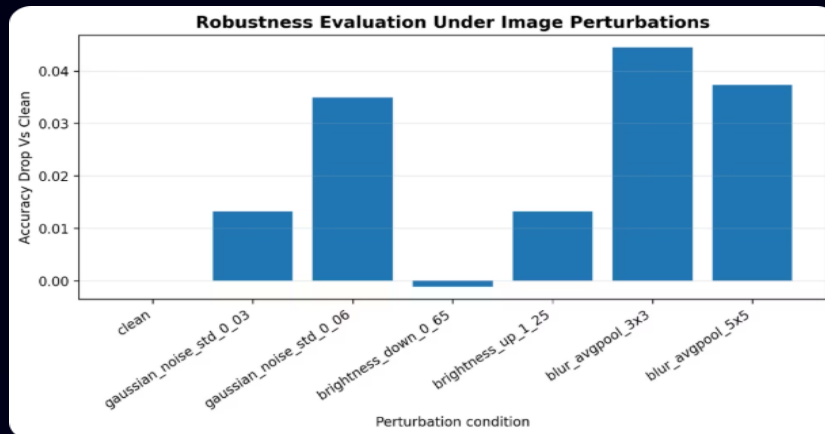
Interpretation

Grad-CAM highlighted image regions associated with the model's predicted class. Correct and misclassified examples showed different activation patterns, but these maps were interpreted only as supporting post-hoc visual evidence—not as proof of causal understanding.

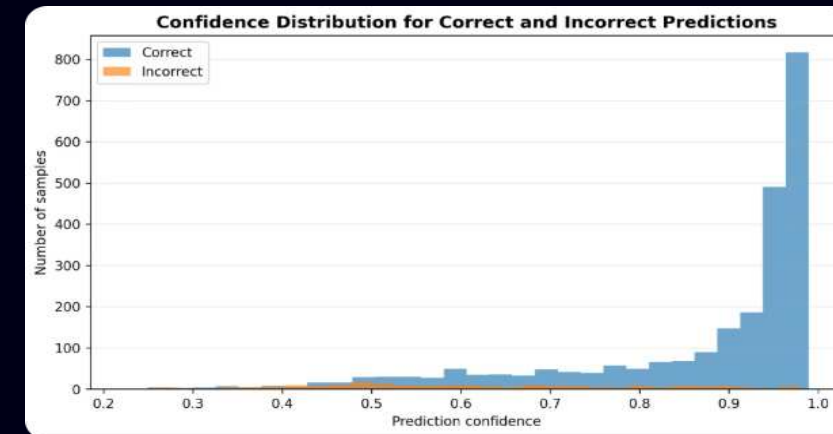
❏ **Key Point** — Grad-CAM provides supporting visual evidence about regions associated with the prediction; it does not prove causal understanding or validate the prediction by itself.

Robustness and Confidence Analysis for Galaxy10

Controlled perturbations and confidence behaviour beyond held-out accuracy.



Robustness under controlled perturbations.



Confidence distributions for correct and incorrect predictions.

Interpretation

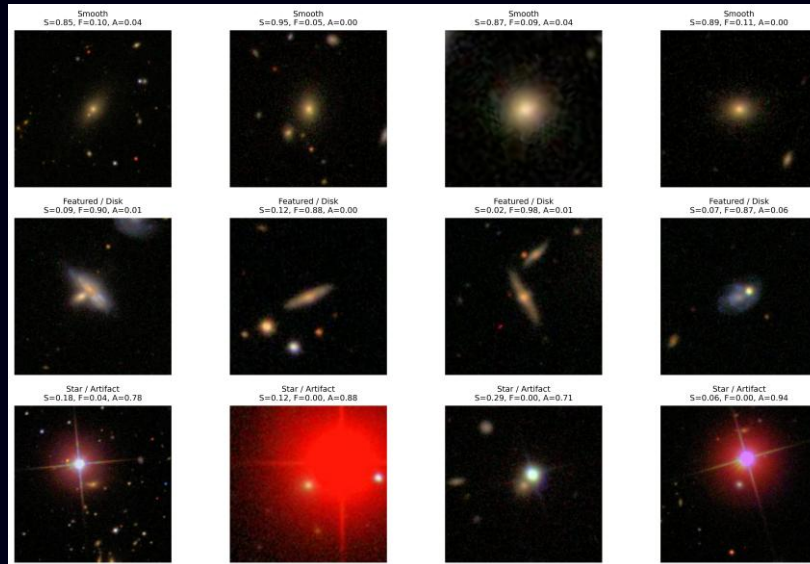
Gaussian noise and blur produced the clearest accuracy degradation, while brightness changes had smaller effects. Correct predictions were generally concentrated at higher confidence, whereas incorrect predictions were more dispersed and tended to show lower confidence. These results provide supporting uncertainty-related evidence, not formal calibration or Bayesian uncertainty estimation.

Key Point – Controlled perturbations and confidence distributions identify conditions associated with reduced model dependability; they do not establish complete robustness or formally calibrated uncertainty.

High-Confidence Smooth vs Featured Classification

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

99.17%
Held-out accuracy

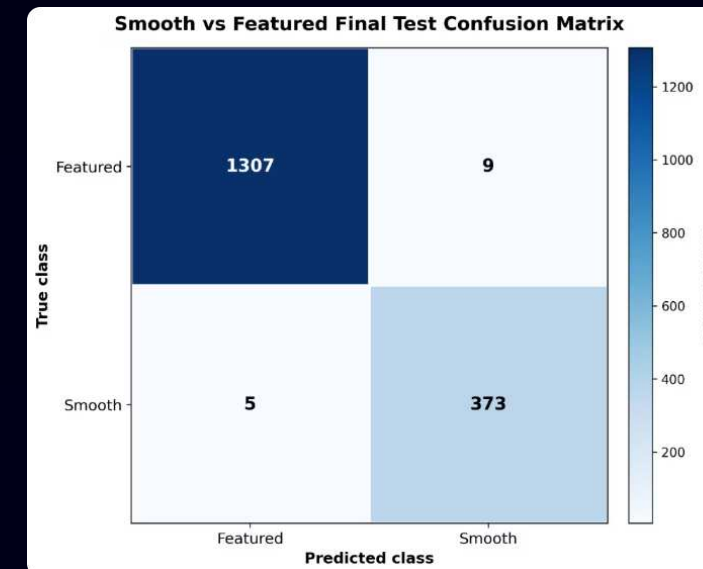


High-agreement first-gate examples; the binary task retained Smooth and Featured/Disk cases only.

Interpretation

The result demonstrates strong held-out routing performance on a high-agreement binary subset. It must not be compared directly with the full ten-class Galaxy10 task.

0.9881
Macro F1

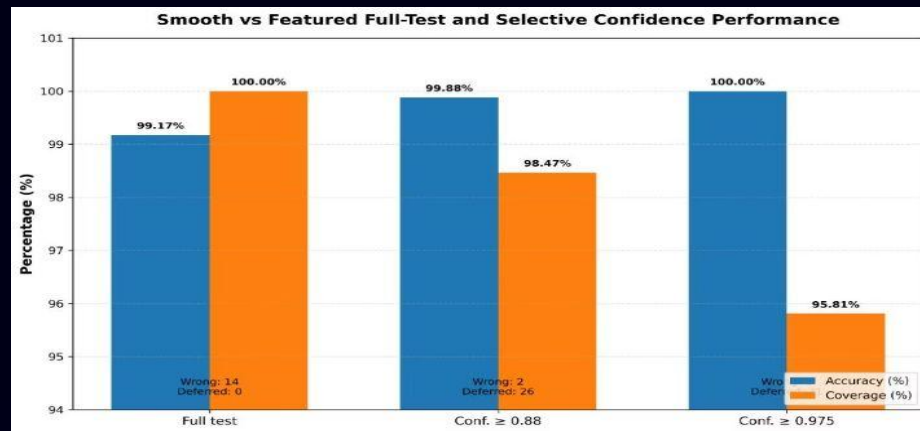


Official held-out confusion matrix.

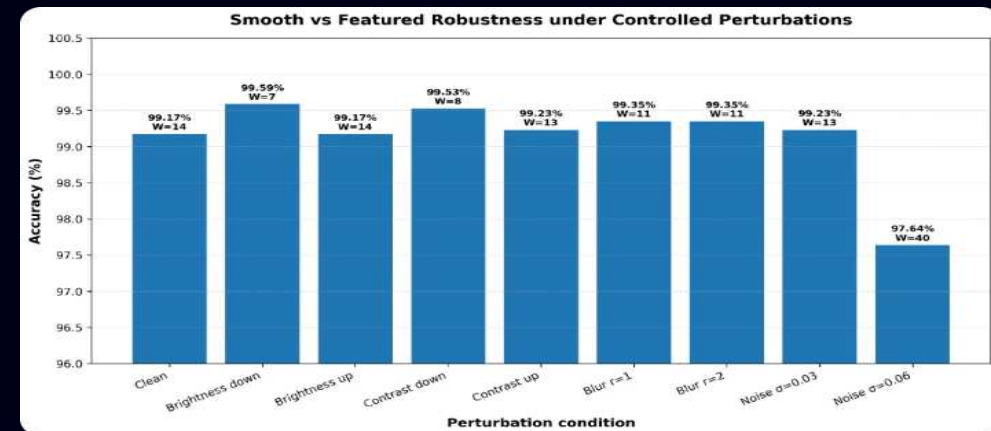
Key Point — The 99.17% result applies only to high-agreement Smooth-versus-Featured routing—not to the full ten-class morphology problem.

Smooth vs Featured: Errors, Confidence, and Robustness

The binary result was evaluated beyond one accuracy number.



Full-test and selective-confidence performance.



Robustness under controlled perturbations.

What the additional analysis shows

Raising the confidence threshold increased selective accuracy from 99.17% to 100%, while coverage decreased from 100% to 95.81%. Most moderate perturbations caused small changes, but the strongest tested noise condition reduced accuracy to 97.64%. The classifier is strong on the selected binary label definition, but uncertain and degraded cases should not be forced into a confident decision.

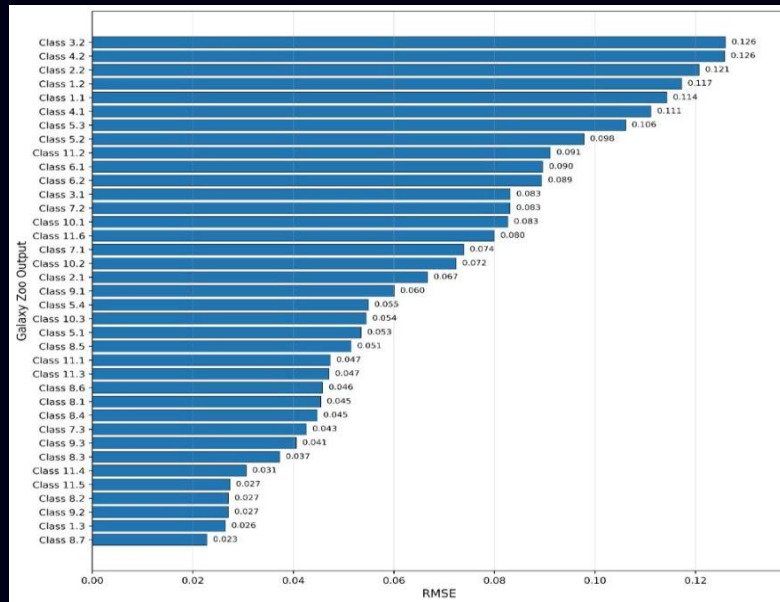
- ❏ **Key Point** — Selective confidence is useful as a routing rule because it allows uncertain cases to be deferred rather than forced into a high-confidence decision.

Galaxy Zoo 37-Output Vote-Fraction Prediction

Continuous morphology prediction on the internal held-out test.

0.074912

Held-out RMSE



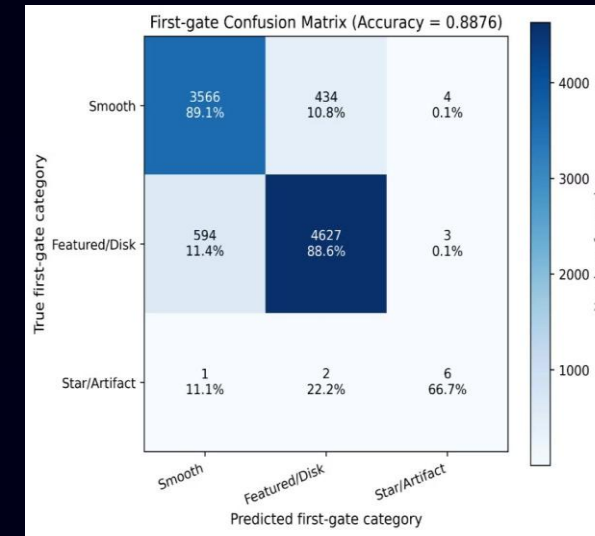
Per-output RMSE across the 37 vote-fraction targets.

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 the primary metrics because the 37 targets are continuous vote fractions, and lower values indicate better prediction. Seven-view TTA improved validation RMSE from 0.07608 to 0.07469 before the selected system was evaluated on the internal held-out test.

Qualitative Error Analysis for Vote Fractions

Representative lowest-error and highest-error predictions.



Representative internal held-out cases ranked by sample-level RMSE across the 37 vote-fraction outputs.

True and predicted labels indicate the dominant first-gate category; RMSE is calculated across all 37 vote-fraction outputs.

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.

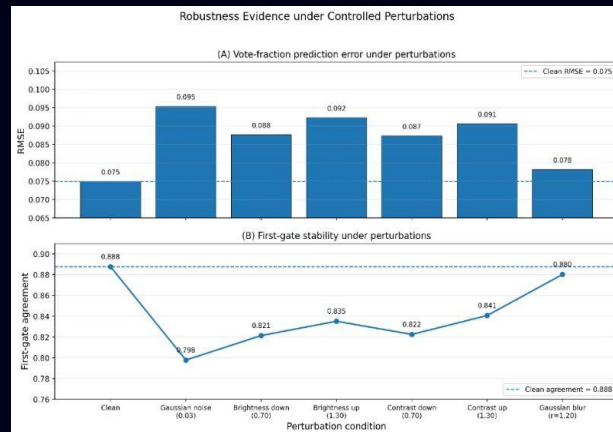
Why this matters

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

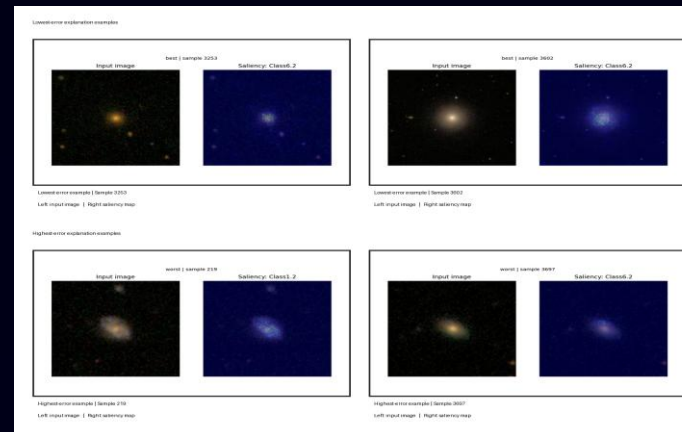
Key Point — The model is evaluated at both aggregate and sample levels; qualitative examples show what a low or high vote-fraction error means in practice.

Framework-Level Reliability Evaluation

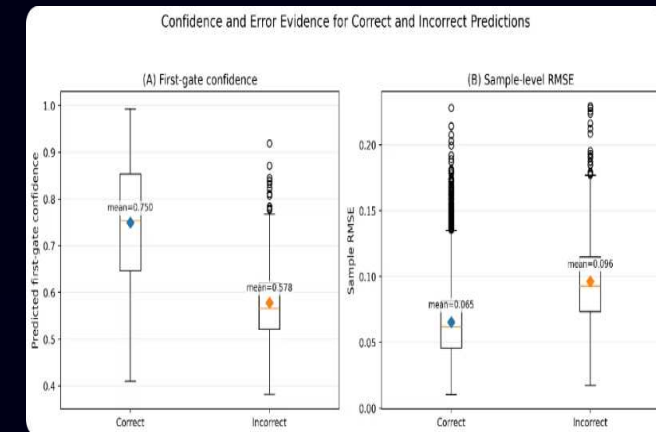
Robustness, confidence, error behaviour, and visual explanation considered together.



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 cases can be flagged at inference, while higher-error cases are identified retrospectively on labelled evaluation data.

Visual explanation finding

Grad-CAM for Galaxy10 and saliency maps for Galaxy Zoo support post-hoc visual inspection, but neither method guarantees correctness or proves causal understanding.

Key Point — Framework-level reliability is assessed through combined evidence from robustness, confidence/error behaviour, and visual explanation; no single score or analysis 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 99.17% 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 the original winning Kaggle benchmark of 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, 99.17% 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

- A task-specific deep learning framework can support galaxy morphology classification and continuous vote-fraction prediction.
- Predictive performance must be complemented with confidence, robustness, vote-fraction modelling, and visual explanation evidence.

Main Contributions

- Structured model-family evaluation
- Confidence-aware routing and selective deferral
- Robustness and confidence-related 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 expert review.

Future Work

- Cross-survey validation
- Confidence calibration
- Probabilistic and Bayesian uncertainty
- Evaluation under 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 any single model is universally best.