Galaxy Morphology Classification with Confidence-Filtered Labels and Robustness-Oriented Modeling

Tiffany Degbotse
Duke University, Pratt School of Engineering
tiffany.degbotse@duke.edu

Sharmil Nanjappa
Duke University, Pratt School of Engineering
sharmilnanjappa.kallichanda@duke.edu

Abstract

Galaxy morphology classification is a foundational task in observational astronomy that supports large-scale studies of galaxy formation and evolution. However, morphology labels derived from citizen science efforts such as Galaxy Zoo are inherently probabilistic and often ambiguous. In this work, we study binary galaxy morphology classification (elliptical vs. spiral) using images from the Galaxy Zoo dataset, with an emphasis on robustness, calibration, and methodological transparency.

We evaluate three modeling paradigms: a naive majority-class baseline, a classical machine learning pipeline using hand-crafted image features with hyperparameter tuning, and deep learning models based on transfer learning with convolutional neural networks. We conduct a focused robustness experiment comparing baseline fine-tuning against astronomy-motivated data augmentation. Additionally, we investigate the impact of confidence-based label filtering, revising an initial strict threshold (> 0.8) to a more permissive threshold (> 0.7) after empirical analysis revealed systematic bias toward easily classified elliptical galaxies.

Our results demonstrate that augmentation-driven robustness substantially reduces overfitting and improves generalization behavior without architectural changes. More broadly, this work highlights the importance of data-centric modeling choices, uncertainty-aware labeling, and careful evaluation when deploying vision models in scientific contexts.

Keywords: Computer Vision, Galaxy Morphology, Robustness, Calibration

1 Problem Statement

Modern astronomical surveys produce massive volumes of galaxy images, rendering manual morphology classification infeasible. Automated galaxy classification systems must therefore be both accurate and reliable under real-world observational variability. Unlike many vision benchmarks, galaxy morphology is not a strictly well-defined concept: transitional forms, low signal-to-noise images, and edge-on spirals routinely challenge even expert annotators.

The objective of this project is to develop and critically evaluate an automated system that classifies galaxies as either elliptical or spiral, while explicitly probing how modeling decisions, particularly label confidence thresholds and data augmentation, affect generalization and robustness.

2 Data Sources and Label Uncertainty

We use the Galaxy Zoo: The Galaxy Challenge dataset, accessed via the Kaggle API. Labels are derived from crowd-sourced volunteer votes aggregated into probabilistic scores following a decisiontree annotation process. These probabilities reflect aggregated human agreement under observational uncertainty, rather than definitive ground-truth morphology.

2.1 Dataset Construction and Splits

To ensure controlled comparison across modeling paradigms and to avoid class imbalance effects, we constructed a balanced subset of the Galaxy Zoo dataset. From the high-confidence labeled samples, we randomly sampled 2,000 elliptical galaxies and 2,000 spiral galaxies, yielding a total dataset of 4,000 images. Sampling was performed using a fixed random seed to ensure reproducibility.

The dataset was split using stratified sampling to preserve class balance across all subsets. First, 15% of the data was held out as a test set. The remaining 85% was then split into training and validation sets, with 20% allocated for validation. This resulted in the following split sizes:

- Training set: 2,720 images (1,360 elliptical, 1,360 spiral)
- Validation set: 680 images (340 elliptical, 340 spiral)
- Test set: 600 images (300 elliptical, 300 spiral)

All classical and deep learning models were trained, validated, and evaluated using the same fixed data splits to ensure fair and consistent comparison across methods.

2.2 Initial Confidence Thresholding

In early experiments, we retained only samples with morphology confidence greater than 0.8. While this reduced label noise, it introduced an unintended effect:

- A strict confidence threshold (> 0.8) simplified the classification task by removing visually ambiguous samples, which accelerated convergence but increased overfitting and degraded generalization to realistic morphology distributions.

2.3 Revised Threshold and Rationale

To better reflect the true morphology distribution, we relaxed the confidence threshold to 0.7. This change reintroduced visually challenging examples, including:

- Faint or low-contrast spirals
- Edge-on spiral galaxies resembling ellipticals
- Transitional or weakly structured morphologies

This adjustment improved the representativeness of the dataset and reduced artificial separability, at the cost of increased label ambiguity, a tradeoff we argue is scientifically appropriate.

3 Related Work

Early work on galaxy morphology classification relied on hand-crafted image features capturing shape, texture, and intensity statistics, typically combined with classical classifiers such as support vector machines [1, 2]. These approaches demonstrated that automated morphology classification was feasible, but their performance was limited by the expressiveness of manually designed features.

More recent studies have shown that convolutional neural networks significantly outperform feature-based pipelines on large-scale galaxy morphology datasets, particularly when trained using transfer learning from natural image corpora [3]. These advances have enabled accurate and scalable automated classification, and have become the dominant paradigm in astronomical image analysis.

However, comparatively less emphasis has been placed on the treatment of label uncertainty, calibration, and robustness in citizen-science-derived datasets such as Galaxy Zoo, where annotations reflect aggregated human agreement rather than definitive ground-truth morphology [4, 5]. Our work builds on prior deep learning approaches by empirically examining how confidence-based label filtering and data augmentation affect overfitting and generalization behavior under realistic annotation uncertainty.

4 Evaluation Strategy and Metrics

We use classification accuracy as the primary metric due to the balanced binary setup. Confusion matrices are reported to analyze class-specific behavior. To diagnose overfitting and generalization, we track training and validation loss across epochs. Importantly, we interpret accuracy in conjunction with loss behavior, recognizing that high accuracy alone may mask overconfidence or poor calibration.

4.1 Metric Selection and Justification

We select classification accuracy as the primary evaluation metric due to the balanced nature of the dataset and the symmetric importance of both galaxy classes. Because the dataset is explicitly constructed with equal numbers of elliptical and spiral galaxies, accuracy provides a clear and interpretable measure of overall model performance without being biased toward a dominant class.

Confusion matrices are reported alongside accuracy to expose class-specific error patterns that are not captured by aggregate metrics alone. In particular, confusion matrices enable analysis of false negatives and false positives for each morphology class, which is critical in astronomical contexts where different error types may have distinct scientific consequences (e.g., misidentifying spiral structure versus smooth morphology).

We intentionally do not prioritize precision–recall metrics or ROC–AUC in this work for two reasons. First, the balanced dataset reduces the risk of misleading accuracy inflation, making accuracy an appropriate primary metric. Second, morphology labels in Galaxy Zoo reflect probabilistic human agreement rather than definitive ground truth. As a result, extremely confident probability-based metrics may overstate model certainty in inherently ambiguous cases.

Instead, we interpret accuracy in conjunction with training and validation loss trajectories to diagnose overfitting and generalization behavior. This combined evaluation strategy emphasizes robustness and reliability rather than purely maximizing a single scalar metric, aligning with the scientific goals of morphology classification under observational uncertainty.

5 Modeling Pipelines

5.1 Naive Baseline

A majority-class classifier predicts the most frequent morphology in the training data. This baseline establishes a lower-bound reference.

5.2 Classical Machine Learning Pipeline

We extract hand-crafted image features including intensity statistics, Hu moments, and gray-level co-occurrence matrix (GLCM) texture features. All features are standardized prior to classification.

We train a support vector machine (SVM) with a radial basis function (RBF) kernel. Hyperparameters are selected via grid search with 5-fold cross-validation over the following ranges:

- Regularization strength $C \in \{0.1, 1, 10, 100\}$
- Kernel coefficient $\gamma \in \{\texttt{scale}, 0.01, 0.001\}$

The optimal configuration, selected based on cross-validated accuracy, was $C = 100$ and $\gamma = \texttt{scale}$. This tuned model is used for all reported classical machine learning results.

5.3 Deep Learning Models

We adopt a ResNet-18 architecture pretrained on ImageNet and fine-tune it for galaxy morphology classification.

We evaluate the following variants:

- A baseline fine-tuned model without data augmentation.
- A robust fine-tuned model trained with astronomy-motivated data augmentation.

Learning Rate Fine-Tuning. As part of the fine-tuning process, we empirically evaluated multiple learning rates (10^{-4}, 10^{-5}, and 3×10^{-6}) to assess training stability and validation performance. While all learning rates achieved high validation accuracy, intermediate values (10^{-5}) consistently produced smoother convergence and more stable validation loss trajectories, particularly when combined with data augmentation. Extremely small learning rates led to slower convergence and increased sensitivity to validation noise, while larger learning rates occasionally introduced minor oscillations in validation loss.

Based on this empirical analysis, we selected a learning rate of 10^{-4} for all reported deep learning results.

6 Focused Experiment: Robustness via Data Augmentation

6.1 Experimental Motivation

Galaxy images exhibit rotational invariance, brightness variability, and atmospheric noise. We hypothesize that incorporating these invariances through data augmentation will reduce overfitting and improve generalization.

6.2 Experimental Design

All models share the same architecture, optimizer, learning rate, and data splits. The only difference is the use of augmentation during training:

- Random rotations
- Brightness and contrast jitter
- Gaussian blur

6.3 Training Dynamics: Baseline vs. Robust Models

We analyze training and validation loss curves to assess overfitting and generalization behavior for baseline and robust models under two different label-confidence thresholds (0.8 and 0.7).

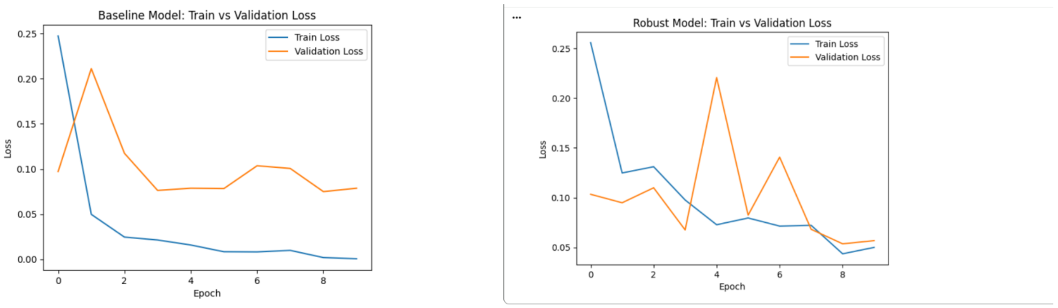

Figure 1: Training and validation loss curves under a high-confidence labeling threshold (> 0.8). **Left:** Baseline model trained without data augmentation, exhibiting rapid convergence and an increasing train–validation gap, indicative of early overfitting on the simplified dataset. **Right:** Robust model trained with data augmentation. Although the robust model exhibits early validation loss spikes, these fluctuations diminish over training and do not lead to a persistent increase in the train–validation gap, indicating more stable generalization despite the simplified nature of the high-confidence labeling.Validation loss spikes observed in the robust model(>0.8) are indicative of stochasticity introduced by augmentation, increased variance due to a smaller validation set and the sensitivity of cross-entropy loss to prediction confidence.

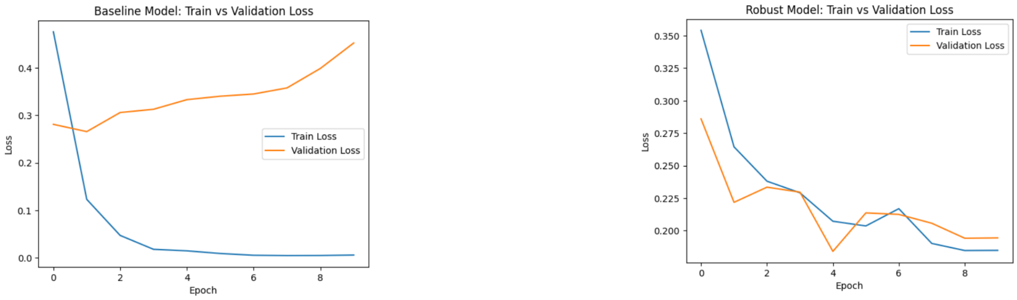

Figure 2: Training and validation loss curves under a relaxed confidence threshold (> 0.7). **Left:** Baseline model trained without augmentation, exhibiting increased overfitting as ambiguous and transitional morphologies are introduced. **Right:** Robust model trained with augmentation, maintaining a smaller train–validation gap despite increased dataset difficulty, indicating improved generalization.

> *Data augmentation substantially reduced overfitting. Unlike the baseline model, the robust model maintains a small and stable gap between training and validation loss across epochs, indicating improved generalization.*

7 Results

Table 1: Test accuracy comparison across modeling paradigms. For deep learning models, results are reported using the relaxed confidence threshold (> 0.7), which provides a more representative evaluation of realistic galaxy morphologies.

Model	**Test Accuracy**
Naive majority-class baseline	0.500
Classical ML (SVM with hand-crafted features)	0.688
Baseline fine-tuned CNN (no augmentation)	0.903
Robust fine-tuned CNN (augmentation, > 0.7 threshold)	0.918

Table 1 summarizes test accuracy across all evaluated models. The naive majority-class baseline performs at chance level, while the classical SVM improves performance but remains substantially below deep learning approaches. Both fine-tuned CNN models achieve high accuracy; however, the robust fine-tuned CNN attains the highest test accuracy (91.8%), slightly outperforming the baseline fine-tuned CNN (90.3%). All deep learning results are reported using the relaxed confidence threshold (> 0.7), which yields a more representative evaluation by including ambiguous and transitional galaxy morphologies.

Table 2: Per-class performance with explicit confusion-matrix counts. Class accuracy is reported as $TP/(TP + FN)$.

Model	**Class**	**TP**	**FN**	**FP**	**TN**	**Class Accuracy**
Classical ML (SVM)	Elliptical	195	105	82	218	0.65
	Spiral	218	82	105	195	0.73
Baseline CNN	Elliptical	275	25	33	267	0.917
	Spiral	267	33	25	275	0.890
Robust CNN	Elliptical	281	19	30	270	0.937
	Spiral	270	30	19	281	0.900

7.1 Per-Class Performance Analysis

Before interpreting per-class confusion-matrix statistics, we first contextualize the naive majority-class baseline. The naive model always predicts the most frequent class in the training data. Because the dataset is class-balanced, this strategy achieves approximately 50% accuracy by construction. However, it provides no meaningful class-specific insight: one class attains perfect recall while the other attains zero recall.

Classical Machine Learning (SVM). The classical SVM model exhibits asymmetric per-class performance. Spiral galaxies are classified more accurately (73%) than elliptical galaxies (65%). This reflects the reliance of hand-crafted features on visible texture and structural cues, which are often more pronounced in spiral galaxies. The relatively high number of false negatives for ellipticals indicates difficulty capturing subtle shape information, particularly for smooth or low-contrast morphologies.

Baseline Fine-Tuned CNN (> 0.7 **Threshold).** The baseline CNN achieves strong and balanced performance, with class accuracies of 91.7% for ellipticals and 89% for spirals. Compared to the classical model, both false positives and false negatives are substantially reduced. However, training dynamics reveal a larger train–validation gap, indicating overfitting despite high test accuracy.

Robust Fine-Tuned CNN (> 0.7 **Threshold).** The robust CNN achieves the strongest overall performance, with 93.7% accuracy on ellipticals and 90% on spirals.The robust model demonstrates improved stability and generalization behavior under increased label ambiguity.

Summary. Overall, deep learning models substantially outperform classical approaches. While the baseline CNN achieves high accuracy, the robust CNN maintains competitive performance while exhibiting improved generalization and reliability under a more representative labeling regime.

8 Error Analysis

To provide a concrete and defensible error analysis, we examine mispredictions from both the baseline fine-tuned CNN and the robust fine-tuned CNN on the validation set. Representative failure cases for each model are shown in Figure **??**. Using these examples, we identify five recurring misprediction types, analyze their underlying causes, and propose targeted mitigation strategies grounded in the current modeling pipeline.

8.1 Case-Based Analysis of Mispredictions

Case 1: Low-Contrast Spiral Misclassified as Elliptical (Baseline and Robust). Observation. A spiral galaxy with weak arm contrast is classified as elliptical by both models. **Root Cause.** Low signal-to-noise suppresses spiral texture cues, causing the galaxy to resemble a smooth elliptical profile in pixel space. **Mitigation.** Incorporate contrast-aware preprocessing and additional low-exposure augmentations to improve robustness to observational noise.

Case 2: Edge-On Spiral Misclassified as Elliptical (Baseline and Robust). Observation. An edge-on spiral galaxy is consistently predicted as elliptical. **Root Cause.** Spiral arms are not visible under extreme viewing angles, and the projected 2D appearance closely matches elliptical morphology. **Mitigation.** Introduce orientation-aware modeling via targeted augmentations or auxiliary geometric features such as axial ratio estimates.

Case 3: Textured Elliptical Misclassified as Spiral (Baseline). Observation. A bright elliptical galaxy with internal brightness variations is misclassified as spiral by the baseline model but correctly classified by the robust model. **Root Cause.** The baseline model overfits to local texture artifacts and illumination gradients, interpreting them as spiral structure. **Mitigation.** Strengthen regularization and expand illumination-based augmentations, as partially achieved by the robust training pipeline.

Case 4: Small Angular-Size Spiral Misclassified as Elliptical (Robust). Observation. A distant spiral galaxy with small apparent size is misclassified as elliptical by the robust model. **Root Cause.** Image resizing and downsampling eliminate fine-grained spiral features, particularly for galaxies occupying few pixels. **Mitigation.** Employ multi-scale feature extraction or adaptive resolution strategies that preserve high-frequency structure for small objects.

Case 5: Ambiguous or Transitional Morphology Misclassified by Both Models. Observation. A galaxy exhibiting weak or transitional morphology is inconsistently classified across models. **Root Cause.** Forced binary labeling does not capture the intrinsic ambiguity present in probabilistic Galaxy Zoo annotations. **Mitigation.** Extend the framework to include an explicit "ambiguous" class or adopt selective classification mechanisms that allow abstention under low confidence.

[t]0.48

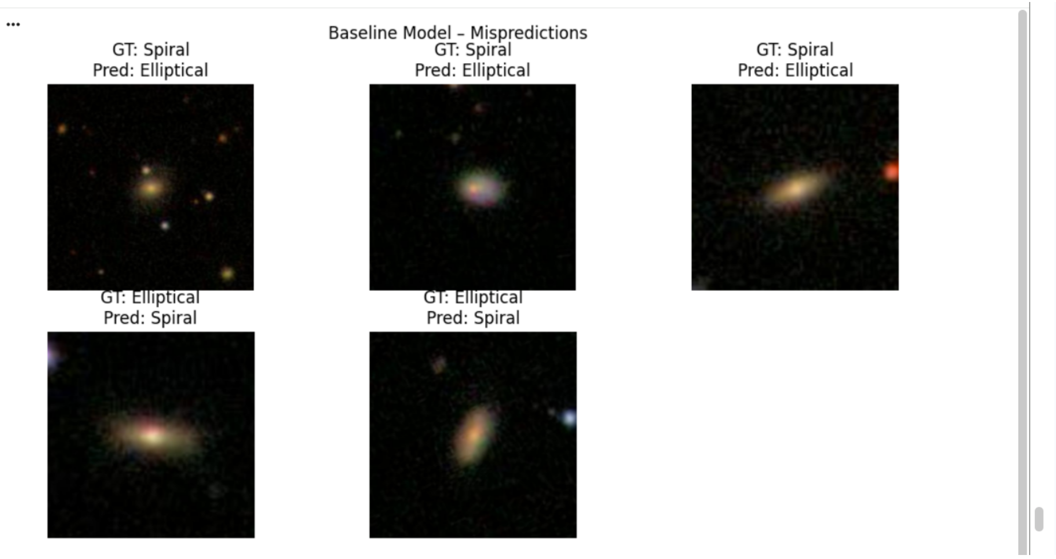

Figure 3: Baseline fine-tuned CNN mispredictions. The model frequently misclassifies low-contrast and edge-on spiral galaxies as ellipticals, and occasionally misinterprets texture artifacts in ellipticals as spiral structure.

[t]0.48

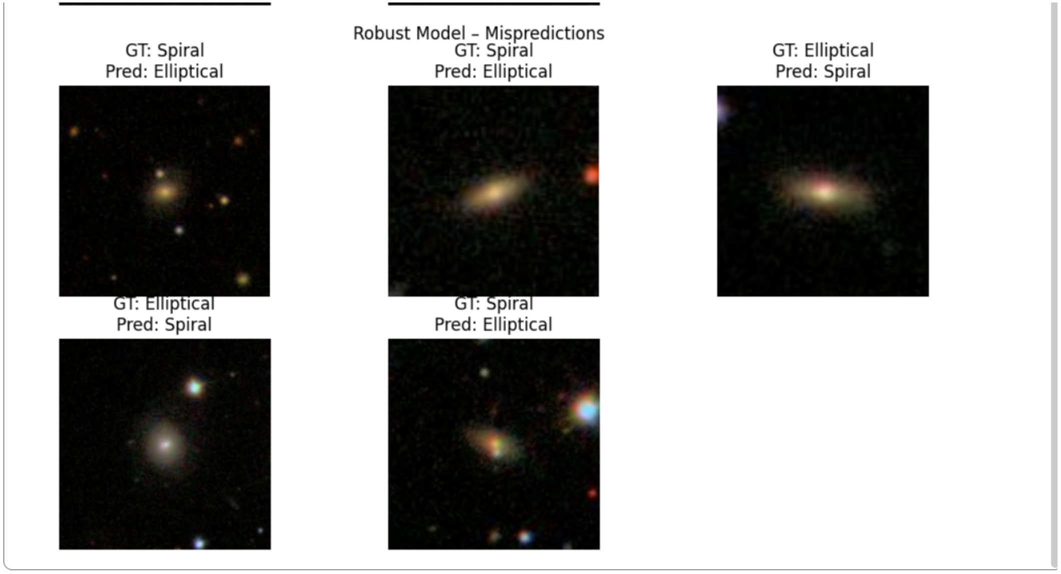

Figure 4: Robust fine-tuned CNN mispredictions. While augmentation reduces texture-driven errors, failures persist for small angular-size galaxies and inherently ambiguous morphologies.

8.2 Summary

Across all five cases, mispredictions arise from a combination of observational limitations, projection effects, and irreducible label ambiguity. While data augmentation substantially improves robustness and reduces overfitting, it cannot fully eliminate errors rooted in intrinsic uncertainty. These findings motivate future work on uncertainty-aware modeling and richer morphological taxonomies.

9 Discussion and Methodological Contributions

This work emphasizes methodology over architecture. Our primary contributions include:

- A confidence-filtered labeling strategy that acknowledges annotation uncertainty
- An empirical analysis of how label thresholds affect overfitting
- A robustness study demonstrating the value of astronomy-motivated augmentation
- A comparative evaluation across naive, classical, and deep learning paradigms

10 Conclusions

Transfer learning combined with data-centric robustness techniques yields strong and reliable galaxy morphology classifiers. While architectural advances remain important, our results demonstrate that thoughtful treatment of labels, augmentation, and evaluation can substantially improve model reliability in scientific applications.

11 Future Work

Future directions include multi-class morphology prediction, explicit modeling of ambiguous cases, uncertainty calibration,and integration of astrophysical metadata.

12 Commercial Viability Statement

This system demonstrates strong potential for real-world use in scientific and industrial contexts, but is not yet suitable for fully autonomous commercial deployment without additional safeguards.

The trained models achieve high accuracy under controlled conditions and exhibit improved robustness through data augmentation and confidence-aware labeling. As such, they are well-suited for use as decision-support tools in astronomical workflows, for example:

- Assisting astronomers with large-scale galaxy catalog curation
- Prioritizing candidate galaxies for expert review
- Serving as a pre-filtering stage in observational pipelines

However, several limitations currently prevent direct end-to-end deployment. First, morphology labels in Galaxy Zoo reflect probabilistic human agreement rather than definitive ground truth, introducing irreducible ambiguity that cannot be resolved by a binary classifier alone. Second, the model does not explicitly quantify predictive uncertainty or provide abstention mechanisms for ambiguous cases, which is critical in high-stakes scientific decision-making. Finally, performance degrades for small angular-size galaxies and edge-on spirals, limiting reliability under certain observational conditions.

From a commercial perspective, the most appropriate deployment model is a human-in-the-loop system, where predictions are accompanied by confidence scores or uncertainty estimates and reviewed by domain experts. With further development, including uncertainty calibration, selective classification, and integration of astrophysical metadata, the system could be extended into a reliable component of professional astronomy pipelines or educational platforms. Until such extensions are implemented, the model should be viewed as a robust research prototype rather than a fully autonomous commercial solution.

13 Ethics Statement

The dataset contains no personal data. We explicitly address label uncertainty and avoid overconfident deployment claims, aligning with responsible AI practices in scientific domains.

References

[1] Lior Shamir. Automatic morphological classification of galaxy images. *Monthly Notices of the Royal Astronomical Society*, 399(3):1367–1372, 2009.

[2] Manda Banerji et al. Galaxy zoo: reproducing galaxy morphologies via machine learning. *Monthly Notices of the Royal Astronomical Society*, 406(1):342–353, 2010.

[3] Sander Dieleman, Kyle W. Willett, and Joni Dambre. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly Notices of the Royal Astronomical Society*, 450(2):1441–1459, 2015.

[4] Chris J. Lintott, Kevin Schawinski, Anže Slosar, 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):1179–1189, 2008.

[5] Chris Lintott, Kevin Schawinski, Steven Bamford, et al. Galaxy zoo 1: data release of morphological classifications for nearly 900,000 galaxies. *Monthly Notices of the Royal Astronomical Society*, 410(1):166–178, 2011.