Lab 3

Jorge Sanchez
(Dated: Decemeber 12, 2025)

The objective of this lab was to develop and apply a Convolutional Neural Network (CNN) model for the automated morphological classification of galaxies drawn from the Sloan Digital Sky Survey (SDSS). Training data consisted of detailed, de-biased vote fractions derived from the Galaxy Zoo 2 (GZ2) citizen science project, covering a sample of over 245,000 galaxies across 37 classification labels. We trained and evaluated two architectures—a custom CNN and a pre-trained ResNet-18—against a naive mean-label benchmark, utilizing techniques including image downsizing and data augmentation. The ResNet-18 model proved most effective, achieving a low overall validation Root Mean Squared Error of $L_{RMSE} = 0.0486$ across the 37 labels, thereby demonstrating high predictive accuracy for complex galaxy features. As a primary application, the trained model was used to estimate the galaxy merger fraction (f_{merger}) for the test sample, resulting in $f_{merger} = 0.766$. This value is significantly higher than the $\approx 2\% - 5\%$ major merger fraction cited in [5][1], a difference that can be attributed to the GZ2 project's broad definition of a merger event and specific properties of the SDSS sample.

I. INTRODUCTION

Galaxies can be divided into categories based on their morphology(shapes) and this has been standard practice since e it was first systematically applied by Hubble in 1936. With sorting this galaxies into categories it surprisingly able to produces classifications which broadly correlate with other physical parameters such as the star formation rate or gas fraction. Catalogs of classified galaxies were complied by individuals or small teams of astronomers for most of the twentieth century. This changes however, with the introduction of modern surveys such as the Sloan Digital Sky Survey (SDSS) which contained many hundreds of thousands of galaxies. This meant that the way galaxies were classified before would no longer be practical. Therefore to Lahav et al. in 1995 compared classification from a set of experts who considered a sample of just over 800 galaxies. This was motivated because they wanted to create a training set for neural networks with the aim of being to automate the classification process.Methods to classify galaxies by morphology have been developed in recent studies still rely on indirect indicators like color, concentration, spectral features, or surface brightness to distinguish early-type galaxies from spirals.While these criteria are useful, they introduce potential biases in the sample, meaning the selected galaxies may not accurately reflect true morphological classifications. As a result, comparing findings from different proxy-based samples can be misleading [4].

To avoid this problem, Galaxy Zoo (GZ1) and later Galaxy Zoo 2 (GZ2) were created as attempts to solve this problem by inviting large numbers of people to classify galaxies over the internet(citizen scientist) to provide morphological classifications for nearly one million galaxies drawn from the SDSS. The difference between GZ1 and GZ2 is that GZ1 divided spiral and elliptical systems, whereas GZ2 aimed to demonstrate a much wider variety of morphological features.

As a prerequisite to analyzing the Galaxy Zoo 2 (GZ2) data, a number of galaxies were classified manually on the Zooniverse platform to gain an understanding of the visual classification process. This experience provided immediate insight into the morphological features that our Convolutional Neural Network (CNN) was trained to identify.Two galaxies encountered during this manual classification were selected as examples of the two main morphological that I saw testing out the galaxy zoo myself:

Galaxy with Clear Spiral Structure (figure 1): This galaxy was classified as a clear spiral due to its well-defined, rotating disk structure and distinct spiral arms emerging from a central bulge. The arms were rich in blue, clumpy star-forming regions, a hallmark of a late-type galaxy.

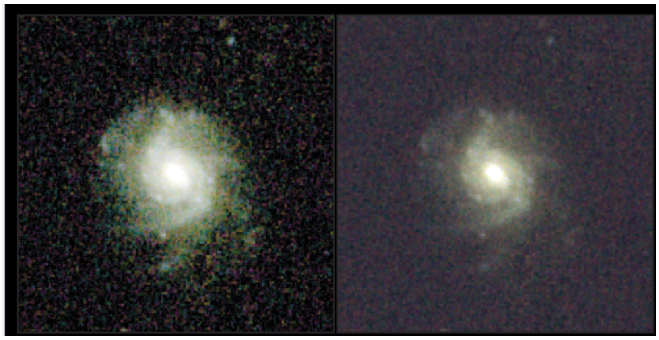

FIG. 1. Example of a clear Spiral Galaxy. This image, taken from the Galaxy Zoo classification interface, illustrates the morphological features of a face-on spiral: a distinct disk, winding arms, and a prominent central bulge.

Morphologically Smooth Galaxy (figure 2): This galaxy was classified as smooth and rounded (an Elliptical galaxy). It was characterized by a smooth, uniform light distribution that faded gradually from a bright, featureless central core. Crucially, it lacked any visible disk, spiral arms, or signs of recent star formation.

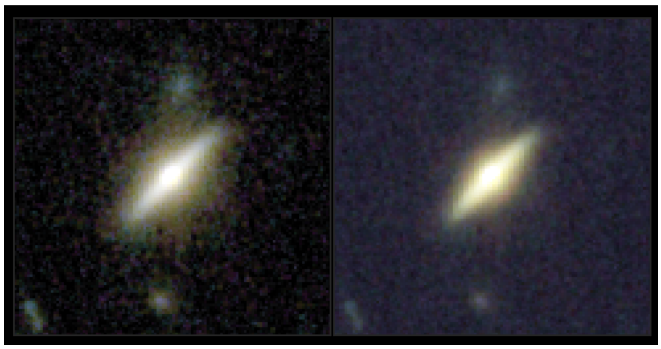

FIG. 2. Example of a Morphologically Smooth Galaxy . This galaxy, classified via the Galaxy Zoo interface, shows the hallmark of an Elliptical galaxy: a smooth, featureless, round-to-oval shape with a uniform brightness gradient from the center.

II. METHODS

A. Data Collection and quality cuts

For GZ2, several cuts were applied to the SDSS Data Release 7 (DR7) Legacy sample because the goal was to include only the nearest, brightest, and largest systems for which fine morphological features can be resolved and classified. GZ2 required a Petrosian half-light magnitude brighter than 17.0 in the r-band, along with a size limit of petroR90_r > 3 arcsec, where petroR90_r is the radius containing 90% of the r-band Petrosian aperture flux.

Galaxies that had a spectroscopic redshift in the DR7 catalog outside the range $0.0005 < z < 0.25$ were removed; however, galaxies without reported redshifts were kept. Finally, objects flagged by the SDSS pipeline as `SATURATED`,`BRIGHT`, or `BLENDED` without an accompanying `NODEBLEND` flag were also removed. This lead to having 245,609 galaxies fitting this criteria which is referred as the 'original' sample.The images that were used are can be described as color composite images that have rectangular grids of 424 x 424 pixels. This does not mean that the angular dimension are the same for all galaxies since the image scaling is variable , designed specifically to encompass the relevant features of the individual galaxy being classified. The scale set can be calculated by using Pixel Scale = (0.02 × petroR90_r arcsec per pixel). The physical scale of each image varies because the pixel scale depends on the galaxy's apparent angular size (Petrosian radius R_{90}). The total angular extent of an image is

$$\Theta = 424 \times (0.02\, R_{90,\mathrm{arcsec}}) = 8.48\, R_{90,\mathrm{arcsec}}. \quad (1)$$

Since the physical Petrosian radius is given by

$$R_{90,\mathrm{phys}} = R_{90,\mathrm{arcsec}}\, D_A, \quad (2)$$

the physical width of an image scales as

$$L_{\mathrm{image}} \approx 8.48\, R_{90,\mathrm{phys}}. \quad (3)$$

Using the adopted cosmology ($H_0 = 71.8\ \mathrm{km\,s^{-1}\,Mpc^{-1}}$, $\Omega_m = 0.273$, $\Omega_\Lambda = 0.727$) and noting that R_{50} is used for debiasing with bins up to $R_{50} \leq 15$ kpc (typically $R_{90} \sim 1.5$–$2\, R_{50}$), the largest galaxies have

$$R_{90,\mathrm{phys}} \sim 25 \text{ kpc}. \quad (4)$$

Thus the physical size represented by an image is approximately

$$L_{\mathrm{image}} \approx 8.48 \times 25 \text{ kpc} \approx 212 \text{ kpc}, \quad (5)$$

so the images correspond to physical scales of order ~ 200 kpc for the largest systems.

B. Classification of Galaxies

Key features classified in GZ2 were; bars, bulges and quantifying their relative strengths, shapes of edge-on disks,tightness and distinction of spiral arms, relative roundness of elliptical galaxies, other odd features such as rings , dust lanes, mergers, disturbed/interacting morphologies , and gravitational lenses. This decision was made by several tasks (questions) which is shown in the figure 3.

The change from GZ1 to Gz2 shifted from simple categorization to a detailed multi step system. Gz1 primarily focused o making only the most basic morphological distinctions separating galaxies into categories of elliptical (early-type) spiral (late-type), and mergers. In contrast GZ2 employed a more complex classification system utilizing a multi step decision tree (11 tasks with 37 possible responses) to measure finer morphological features such as bars, bulges, the shapes of edge on diskls, and the relative strengths of galactic bulges and spiral arms which again can be seen in figure 3.

For most of it duration GZ2 showed images to classifiers by randomly selecting them from the database. However toward the end of the project a prioritization mechanism was implemented to ensure the necessary data completeness.Importantly a volunteer could not choose which galaxy to classify once a classification was complete the next galaxy image was automatically displayed.GZ2 main sample received 44 unique classifications which were made by having GZ2 have an iterative weighting scheme that was applied to users to reduce the influence of those whose votes were inconsistent with the consensus. This scheme involved calculating a users consistency () by comparing their votes to the overall vote fraction for a task, and then applying a weighting function (ω) that down weighted classifiers in the tail of low consistency. Additionally any repeated classifications of the same image by the same user were removed retaining only their votes from the last submission to ensure each vote was treated as an independent measurement.

The primary reason is classification bias where observers find it more difficult to identify finer morphological features in galaxies that are more distant (therefore

Task	Question	Responses	Next
01	*Is the galaxy simply smooth and rounded, with no sign of a disk?*	smooth	07
		features or disk	02
		star or artifact	**end**
02	*Could this be a disk viewed edge-on?*	yes	09
		no	03
03	*Is there a sign of a bar feature through the centre of the galaxy?*	yes	04
		no	04
04	*Is there any sign of a spiral arm pattern?*	yes	10
		no	05
05	*How prominent is the central bulge, compared with the rest of the galaxy?*	no bulge	06
		just noticeable	06
		obvious	06
		dominant	06
06	*Is there anything odd?*	yes	08
		no	**end**
07	*How rounded is it?*	completely round	06
		in between	06
		cigar-shaped	06
08	*Is the odd feature a ring, or is the galaxy disturbed or irregular?*	ring	**end**
		lens or arc	**end**
		disturbed	**end**
		irregular	**end**
		other	**end**
		merger	**end**
		dust lane	**end**
09	*Does the galaxy have a bulge at its centre? If so, what shape?*	rounded	06
		boxy	06
		no bulge	06
10	*How tightly wound do the spiral arms appear?*	tight	11
		medium	11
		loose	11
11	*How many spiral arms are there?*	1	05
		2	05
		3	05
		4	05
		more than four	05
		can't tell	05

FIG. 3. The Galaxy Zoo 2 classification schema is a decision tree with 11 tasks and a maximum of 37 possible classification labels (features) that a citizen scientist can assign to a galaxy image. The classification process branches based on previous responses.[6]

generally smaller and dimmer). This difficulty leads to a systematic decrease in the fraction of votes for features like spiral structure or galactic bars at a higher redshift which mimics a change in observed morphology fractions that is independent of any true evolution in galaxy properties.

Classification bias is the phenomenon where the observed morphological fractions of galaxies appear to change as a function of redshift independent of any true galaxy evolution, This occurs because finer morphological features (spiral structures or galactic bars) are more difficult to identify in galaxies that are smaller and dimmer due to greater distance.To correct this GZ2 employs a method that derives de biased vote fractions by assuming that the true ratio of classification likelihoods for galaxy of a given physical brightness and size should be consistency across different redshift bins. GZ2 calculates a multiplicative correction constant based on the difference between the ratio of measured vote fractions at a given redshift and the ratio found at the lowest presumably unbiased redshift slice. This allows GZ2 to adjust the raw vote fractions for individual galaxies yielding consistent morphology fractions over a range of redshifts.

Although Galaxy Zoo 2 (GZ2) utilizes boolean "yes/no" questions within its multi-step decision tree, the final reported classifications are floats (vote fractions and debiased likelihoods ranging from 0 to 1) because each galaxy is classified independently by a large number of volunteers [26 within its multi-step decision tree, the final reported classifications are floats (vote fractions and debiased likelihoods ranging from 0 to 1) because each galaxy is classified independently by a large number of volunteers. These floats represent the fraction of total votes cast for a specific response for that galaxy, which acts as an estimate of the classification likelihood or the strength of the feature in question. This methodology allows GZ2 to move beyond a single categorical assignment to provide a probabilistic weight for analyzing the entire sample

The most significant differences between GZ2 classifications and other morphological catalogs lie in scale, detail, and specific feature identification. GZ2 provides classifications for more than an order of magnitude more galaxies (over 300,000) than the largest expert-classified catalogs, enabling large-scale statistical studies. While GZ2 achieves good agreement with experts on fundamental traits like T-types and strong bars, it exhibits less confidence in identifying weak, inner, and nuclear bars/rings due to its methodology. Compared to automated methods, GZ2 classifications are necessary because automated systems still suffer drawbacks and benefit from the GZ2 data as a large, detailed training set.

C. Motivation for Classification

Astronomers care about morphological classification because a galaxy's shape encodes its physical state and evolutionary past. Morphology tells us about merger history,star formation, gas, accretion, environment,internal dynamical evolution. Some examples of this is how elliptical show evidence of past major mergers, spirals having quiet histories, ongoing star formation, irregulars having interactions or high-redshift turbulence, Bars have secular evolution and gas inflow, and cluster early-types having environmental transformation [2].

Astronomers systematically search for merging galaxies because mergers drive galaxy evolution, trigger star formation and AGN activity, reshape morphology, and provide essential tests for cosmological simula-

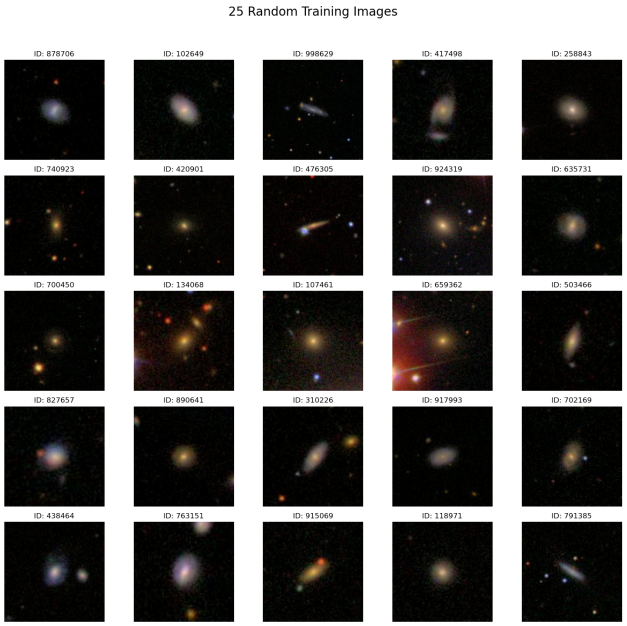

FIG. 4. C25 Random Training Images. A sample of 25 galaxies randomly selected from the GZ2 training dataset, which consists of 61,578 images, illustrating the variety of morphologies available for classification.

tions.Historically, mergers have been identified through morphology,close pairs, kinematic/spectroscopic signatures, and simulation merger trees.

D. Inspecting Data

To start inspecting the data the first step was to download the training set that was provided to us for this lab and in that it was file there was 61578 in the training set with dimensions (124 × 124 pixels). In figure 4 shows 25 random images from the training set to get a picture of what the galaxies look like. Then plotting the distributions of the training shown in figure 5 which can be seen that overall almost none of the distributions are "bell-shaped" (Gaussian) which is completely expected. The columns represent the fraction of votes from citizen scientists where a peak at 0.0 means high agreement (feature is absent),a peak at 1.0 means high agreement (feature is present), and a peak in the middle means it is ambiguous.Now to better illustrate this, 6 (Prototype Image for Each Label) shows the image for which each label has the highest value. From 6, spirals will be easy to classify since they are visually obvious, whereas the bulge shapes will be hard to tell what shape they are visually and therefore will be more difficult for the neural network to classify as well.

To get a better picture of how each label is correlated to each other plotting the correlation matrix which is given in equation **??** which the plot is shown in figure 7 and

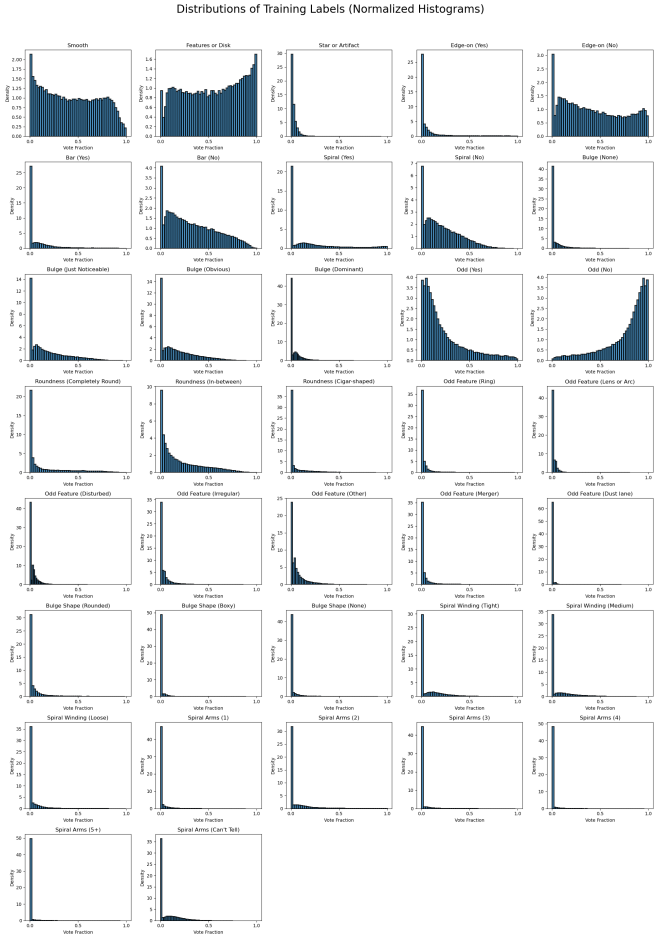

FIG. 5. Distributions of Training Labels (Normalized Histograms). Normalized histograms for each of the 37 classification labels in the training set. A peak at 0.0 indicates a high agreement that a feature is absent; a peak at 1.0 indicates high agreement that the feature is present. A peak in the middle represents an ambiguous or transitional classification.

an interesting thing that we can see is that we can see is that many features are strongly anti-correlated, which is expected. For example, the vote fraction for "smooth" galaxies (Class 1.1) is highly anti-correlated with the vote fraction for "spiral arms" (Class 4.2), meaning a galaxy is highly unlikely to be both smooth and have visible spiral arms..

$$\rho_{ij} = \frac{\langle ij \rangle - \langle i \rangle \langle j \rangle}{\sqrt{\langle i^2 \rangle - \langle i \rangle^2} \sqrt{\langle j^2 \rangle - \langle j \rangle^2}} \tag{6}$$

E. Training the model

Now that we know how the classification work we can know talk about the model that is going to be able to predict classifications (labels) based on images. To train

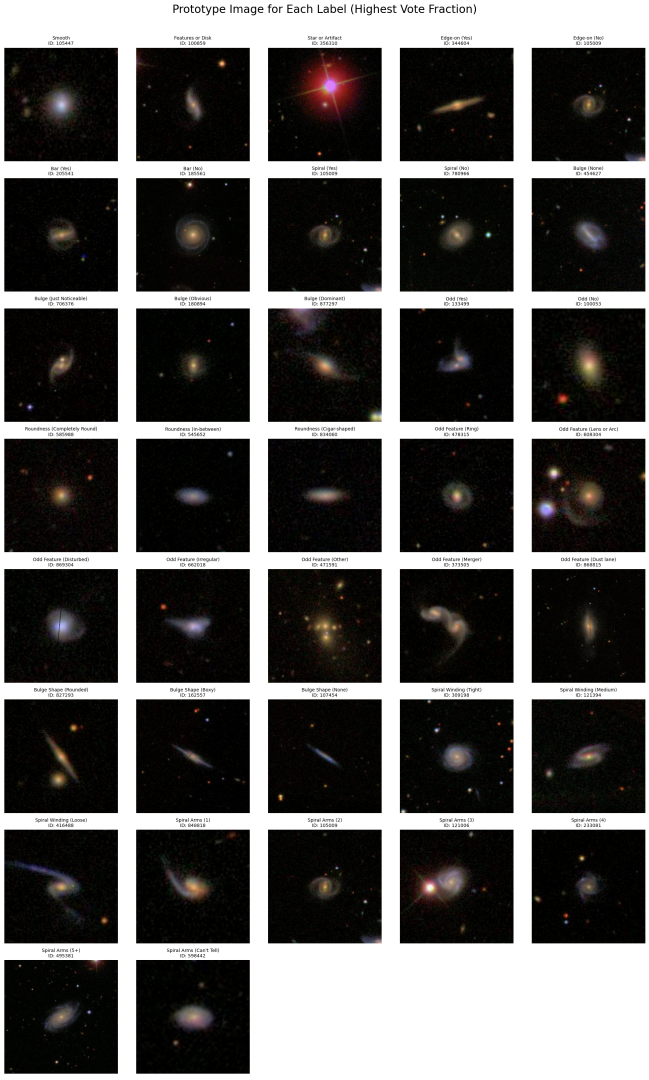

FIG. 6. Images in the training set that is the prototype of that label;i.e.,the image for which the label has the highest value

the model we will have to repeatedly compare predictions for reach image in the training set to its known labels.To address the rest of the questions in Task 9, we note that the visual inspection of the normalized distributions in Figure 7 confirms there are no systematic differences between the training and validation label sets, ensuring the model's performance on the validation set will be a fair measure of its generalization ability. To reduce computational cost.

To reduce computational cost of classification and memory requirements is to downsize the images which was done by cropping the images (removing empty space at the edges) and resampling them to a coarser pixel grid. In figure 8 we can see a comparison of a few images before and after downsizing.

Another way to cost of computational cost we were able to read in the data in batches and for this lab I set

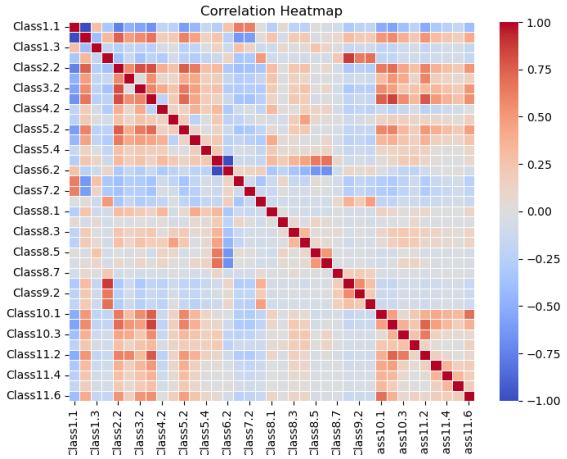

FIG. 7. Correlation Heatmap of GZ2 Classification Labels. The correlation matrix ρ_{ij} for all 37 classification labels in the training set, calculated using Equation 6. High positive or negative correlations (darker colors) indicate features that are frequently voted present or absent together.

the batch size to 128.After downsizing we can now split the images and labels into a training set (containing 80 % of all objects) and a validation set (containing the remaining 20 %). To ensure that the training and validation labels have no systemic differences between them I compared the normalized distributions of them which can be shown in figure 9

Before implementing a neural network we first tried something simpler and that was by guessing the labels for each images as the mean value of the corresponding label in the training set which is done without looking at the image at all. In other words guessing the same label for each image which isn't going to crude guess but provides a benchmark that the neural network model will outperform. Using the equation 8 7 we are able to find the loss function which for our case is the root mean squared error.

$$L_{\text{RMSE}} = \sqrt{\left\langle \left(\overleftrightarrow{\ell}_{\text{true}} - \overleftrightarrow{\ell}_{\text{pred}} \right)^2 \right\rangle} \tag{7}$$

$$L_{\text{RMSE}} = \sqrt{\frac{1}{N_{\text{galaxies}} N_{\text{labels}}} \sum_{i}^{N_{\text{galaxies}}} \sum_{j}^{N_{\text{labels}}} (\ell_{\text{true},ij} - \ell_{\text{pred},ij})^2} \tag{8}$$

Where $\overleftrightarrow{\ell}_{\text{pred}}$ and $\overleftrightarrow{\ell}_{\text{true}}$ are matrices (tensors, in PyTorch parlance) representing the predicted and true labels for each object.For this simple model the $L_{\text{RMSE}} = 0.1638$ for the training set and $L_{\text{RMSE}} = 0.1641$ for the validation set.

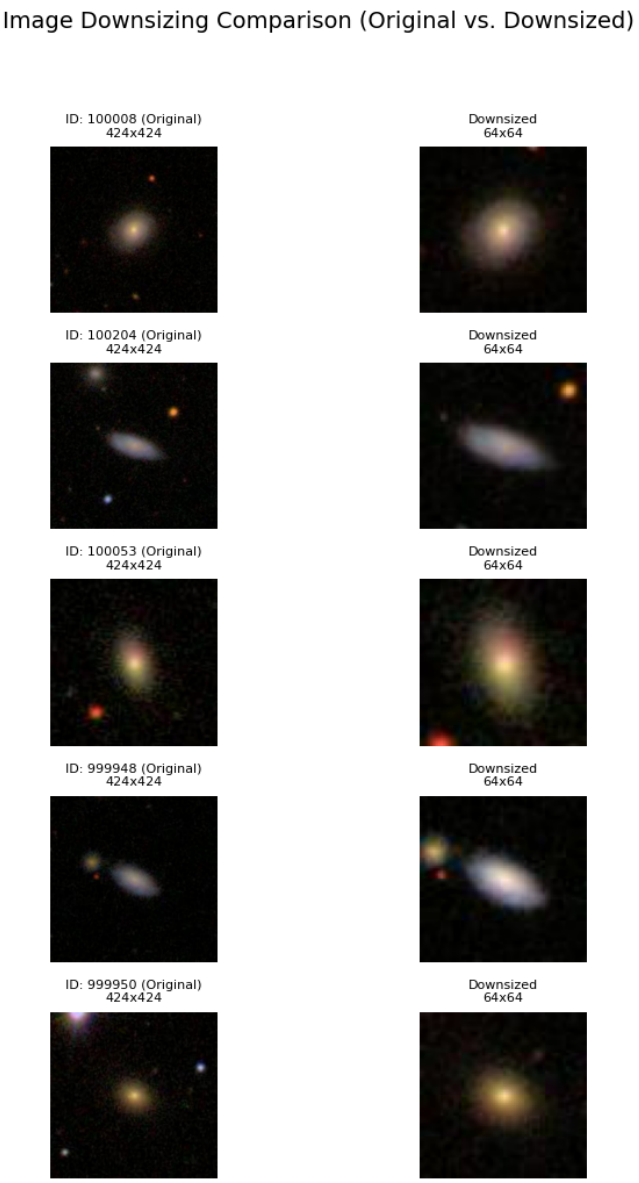

FIG. 8. Image Downsizing Comparison (Original vs. Downsized). Comparison of original 424×424 pixel images (left) with their downsized 64×64 counterparts (right), which was done to reduce computational cost and memory requirements8.

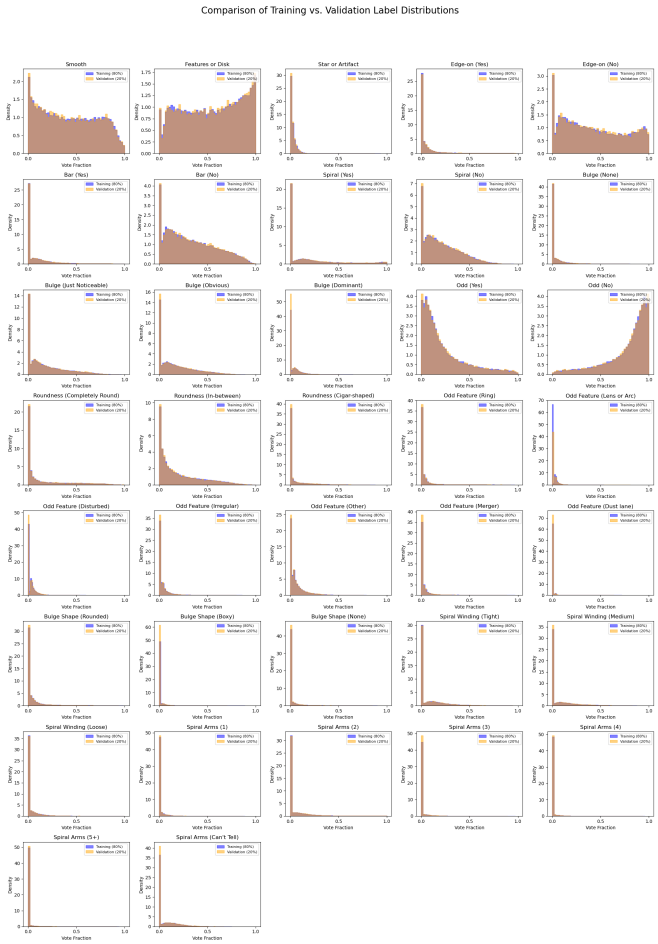

FIG. 9. Normalized Distributions of Training and Validation Labels. Comparison of the normalized label distributions for the training (80%) and validation (20%) sets, confirming no systemic differences between the two sets.

F. ResNet CNN

Before implementing our own convolution al neural network (CNN) we first used an already prebuilt CNN architecture which for this lab is the Residual Network (ResNet).A Resnet is a deep CNN architecture built by stacking "residual blocks". The key feature of a residual block is a shortcut connection (or skip connection) that bypasses one or more layers, feeding the input directly to the output of the block.

The difference from just stacking traditional CNN layers is that traditional deep CNNs suffer from the degradation problem; adding more layers can lead to higher training error, suggesting that the network cannot easily learn the identity mapping needed for new layers to simply pass information through. ResNets solve this by explicitly formulating the layers to learn a residual mapping, $F(x)$, instead of the desired full mapping, $H(x)$. The relationship is defined as $H(x) = F(x) + x$. This shortcut connection makes it easier for the block to learn the identity function ($F(x) = 0$) if needed, enabling the training of much deeper networks without performance degradation [3].

using the 18-layer ResNet on our galaxy images where I modified the base architecture to accommodate the input and output sizes and the fact the the outputs are normalized between 0 and 1. To monitor the performance of the network during training plotting both the training and validation loss as a function of the number of training images the network has seen and this is shown in figure 10.

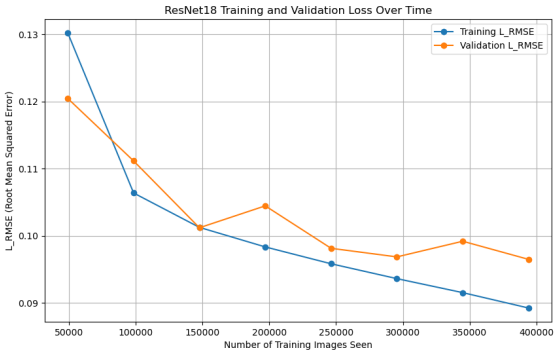

FIG. 10. ResNet18 Training and Validation Loss Over Time. The root mean squared error (L_{RMSE}) is plotted as a function of the number of training images seen, showing the rapid decrease in loss during training. The final $L_{RMSE} = \mathbf{0.1638}$ for the training set and $L_{RMSE} = \mathbf{0.1641}$ for the validation set.

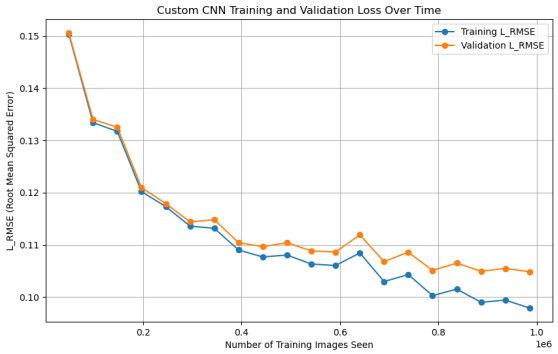

FIG. 11. Custom CNN Training and Validation Loss Over Time. The root mean squared error (L_{RMSE}) for the custom Convolutional Neural Network (CNN) as a function of the number of training images seen.

G. Own CNN

Next I implemented my own CNN to predict labels based on images. I made so it is composed of four sequential convolutional blocks, followed by two dense (fully connected) layers, designed to process 64×64 RGB galaxy images. The architecture consists of four repeating blocks of `Conv2d` $\rightarrow$ `ReLU` $\rightarrow$ `MaxPool2d`, progressively increasing the filter count from $3 \rightarrow 16 \rightarrow 32 \rightarrow 64 \rightarrow 128$. Specifically, the first convolutional layer is defined as `nn.Conv2d`$(\text{in} = 3, \text{out} = 16, \text{kernel\_size} = 5, \text{stride} = 2, \text{padding} = 2)$, while the subsequent `Conv2d` layers use a kernel size of 3 and stride of 1, with 1 pixel of padding. The repeated application of `nn.MaxPool2d`$(\text{kernel\_size} = 2, \text{stride} = 2)$ spatially down samples the 64×64 input to a final feature map of 2×2 across 128 channels. This $128 \times 2 \times 2 = 512$ flattened feature vector is then fed into the first fully connected layer, `nn.Linear`$(512, 512)$, which uses a `ReLU` activation function $f(x) = \max(0, x)$ and is regularized by `nn.Dropout`$(p = 0.5)$. The final layer, `nn.Linear`$(512, 37)$, is followed by the Sigmoid activation function, $\sigma(x) = \frac{1}{1+e^{-x}}$, which bounds the 37 outputs to the range $[0, 1]$, making the model suitable for predicting the independent vote fractions for each galaxy classification label. The end result of this model can be shown in figure 11 of the training and validation loss function as a function of the number of images the model has seen.

H. Optimizations of models

To improve the performance of the network by adjusting the learning rate of the optimizer and using a scheduler (e.g., `torch.optim.lr_scheduler.ReduceLROnPlateau`) to decrease the learning rate time(`lr=0.001`) the validation loss plateaus and this was done with the ResNet model.

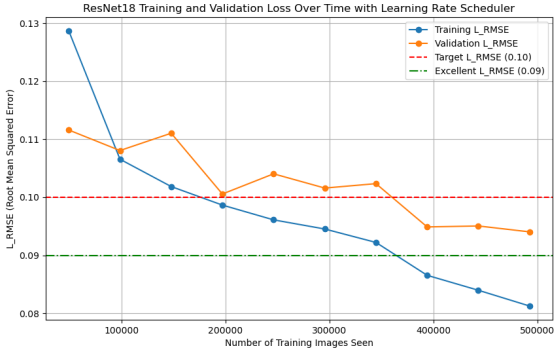

FIG. 12. The training and validation loss curves (Root Mean Squared Error, L_{RMSE}) for the ResNet-18 model as a function of training epoch. The curves converge to a validation loss of $L_{RMSE} = 0.0486$, indicating successful training and generalization.

After training for a while the training loss continues to decrease while the validation loss reaches a plateau which is shown in figure 12 which is evidence of over training. A way to reduce over training is data augmentation which is artificially increasing the size of the training set. This was done by exploiting rotational and reflection symmetry so applying this to our images bu rotating them to a random angle $\theta\epsilon[0.360]$ prior to being cropped and resized. This will make it so that the network will never see the same

image twice even after many epochs of training. With this new optimization shown in figure 13 we can the loss function decrease to below 0.09 which is an excellent loss function value.

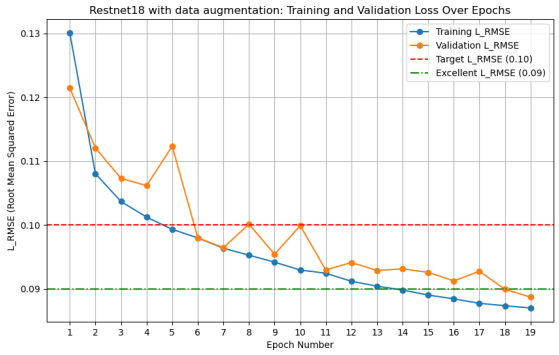

FIG. 13. The training and validation loss curves for the ResNet-18 model with data augmentation applied. Data augmentation helped to regularize the model and minimized the gap between training and validation loss, resulting in a robust final model.

I. Comparing CNN's

Now since we have made different CNN's its a good idea to compare them and see how they differ and the difference of their loss function which is shown in 14.

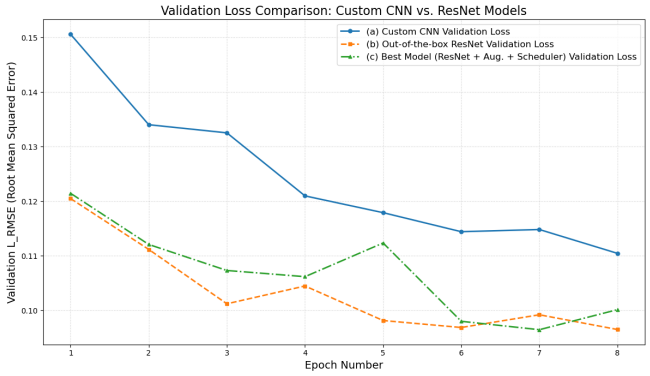

FIG. 14. Comparison of Validation Loss (L_{RMSE}) for Different Neural Network Architectures. This figure compares the validation loss curves for the custom CNN and the ResNet-18 model (with and without enhancements) as a function of training epoch. The comparison illustrates the superior and more stable convergence of the best-performing model.

J. Reliability of model's classification

Now that we have a good sense that our model is working next it to see how reliable it is for classifying glaxies. This was done by comparing the "true" and "predicted"

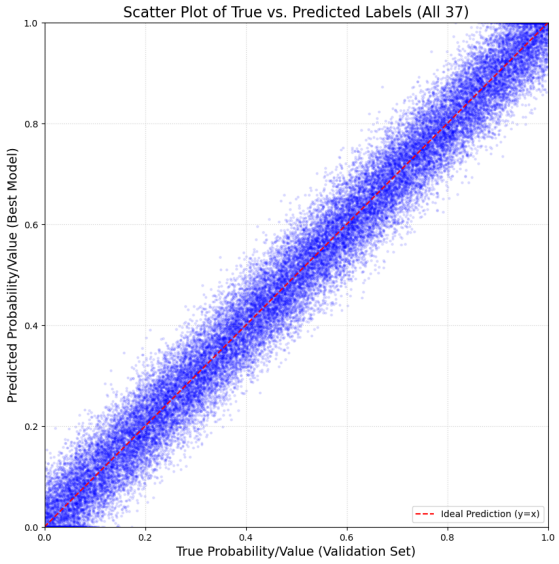

FIG. 15. Scatter Plot of Predicted vs. True Vote Fractions for All 37 GZ2 Labels. The plot compares the model's predicted vote fractions (y-axis) against the corresponding true debiased vote fractions (x-axis) for all 37 labels in the validation set. The tight clustering of data points along the $y = x$ line demonstrates the model's high accuracy and strong predictive power across all classification tasks.

values of all 37 labels forthe validation set, using the best model which was the ResNet 18 with data augmentation and learning rate scheduler. This scatter plot is shown in figure 15 providing a visual assessment of the model's predictive performance. The predicted values are tightly clustered around the ideal prediction line ($y = x$), indicating a strong correlation between the model outputs and the true labels. The symmetry of the scatter about this line suggests minimal systematic bias, with no clear tendency toward overprediction or underprediction. Additionally, the model maintains high accuracy at the extremes of the label range, as evidenced by the dense clustering near values of 0.0 and 1.0. Overall, this behavior demonstrates that the RestNet 18 model is both accurate and robust across the full range of values and for all combined labels.

Now looking at the overall validation root-mean-square error across all 37 labels is $L_{\mathrm{RMSE}} = 0.0486$ (16), indicating a low prediction error that is consistent with the strong agreement observed in the true versus predicted scatter plot. The individual label L_{RMSE} values are highly consistent and cluster closely around this overall mean, demonstrating that the model performs uniformly well across the full set of morphological classifications. The strongest performance is achieved for the *Odd Feature (Disturbed)* label, with a single-label L_{RMSE} of 0.0466, while the weakest performance occurs for the *Spi-*

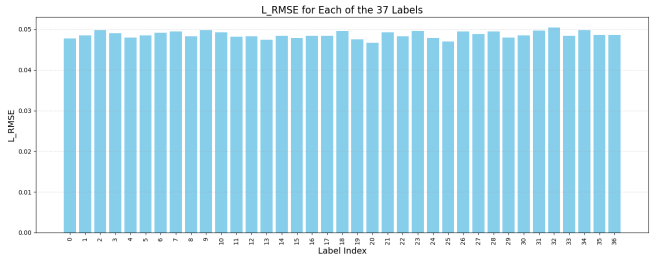

FIG. 16. Root Mean Squared Error (L_{RMSE}) for Each of the 37 Classification Labels. This bar chart displays the individual L_{RMSE} value for each of the 37 Galaxy Zoo 2 classification labels on the validation set. The consistency of the error values across all labels, clustering closely around the overall mean $L_{\mathrm{RMSE}} = 0.0486$, indicates the model's stable and uniform performance across all classification tasks.

ral Arms (2) label, with a value of 0.0505. Notably, the difference between the best- and worst-performing labels is only 0.0039, underscoring the robustness and stability of the model's predictions across all 37 classification tasks.

1. *Classes with Great Performance*

Based on the provided extreme examples, the reliability of the model's classifications can be assessed by comparing the **True** label probability (the ground truth from the validation set) with the **Pred** label probability (the model's output). A strong model is indicated when the predicted probability aligns closely with the true probability, especially for the top-ranked examples.

- **(1) smooth:**
  - **Reliability:** The model performs **very well**. For the majority of the top-ranked predictions, the predicted probability is very high (≥ 0.90) and closely matches the high true probability (≥ 0.90). The examples visually represent typical smooth elliptical galaxies, indicating the model has learned the visual features of this class effectively.
- **(3) edge-on disk:**
  - **Reliability:** The model demonstrates **high reliability**. The predicted probabilities are consistently very high (≥ 0.95), and the true labels are also nearly 1.000. This is an unambiguous morphological class, and the model clearly identifies the elongated, narrow disk structure when viewed edge-on.
- **(7) odd: merger:**
  - **Reliability:** The model shows **high reliability**. The predicted probabilities are consistently high ($\approx$ 0.90 to 0.99) and align well with the high true probabilities. Mergers often exhibit disturbed, asymmetric, or multiple-nucleus morphologies, and the model appears robust in identifying these complex, high-probability examples.
- **(4) odd: ring:**
  - **Reliability:** The model shows **moderate to good reliability**. The model assigns high probabilities (up to ≈ 0.995), but there are cases where the predicted probability is notably lower than the true probability (e.g., True: 0.952 — Pred: 0.719 or True: 0.975 — Pred: 0.595). This suggests that while it correctly identifies the extreme examples, it may be **less confident** for some high-probability true ring galaxies, or that some non-ring examples are mistakenly assigned a high prediction.
- **(6) spiral: 2 arms:**
  - **Reliability:** The model's performance appears to be **moderate**. While the predicted probabilities are the highest for this class (ranging from ≈ 0.60 to 0.94), they are generally lower than the true probabilities. The true labels are often quite high (e.g., 0.829, 0.775, 0.863), but the predicted values often fall short (e.g., 0.508, 0.385, 0.630). This suggests **low confidence** in its predictions for spiral arms, which is a common challenge in galaxy morphology due to viewing angle and the faintness of arms.
- **(5) odd: lens/arc:**
  - **Reliability:** The model exhibits **poor reliability** for this label. The predicted probabilities (ranging from ≈ 0.53 to 0.96) are often significantly lower than the true probabilities (up to 1.000). The visual examples show typical elliptical/disk galaxies, and the faint lens/arc features (gravitational lensing) are **not consistently visible** or are subtle, suggesting the model is likely misclassifying many standard galaxies or is failing to confidently identify the subtle gravitational lensing signature.
- **(2) star/artifact:**
  - **Reliability:** The model shows **poor reliability**. The predicted probabilities are very low for the top extreme examples, especially compared to the true probabilities (e.g., True: 0.775 — Pred: 0.588 and True: 0.876 — Pred: 0.351). The top predicted examples include bright stars, diffraction spikes, or saturated pixels. The low predicted probability suggests

the model is **highly uncertain** about what constitutes a "star/artifact" or that this class is highly imbalanced/ambiguous, causing low model confidence even for visually clear examples.

2. Classes with Poor Performance and Potential Causes

The model performs notably poor for **(2) star/artifact** and **(5) odd: lens/arc**, and shows reduced confidence for **(6) spiral: 2 arms**.

- **Label (2) star/artifact:**
  - **Cause:** The low predicted probabilities suggest the model is struggling to correctly assign high confidence to this class. This is often due to **data imbalance** (artifacts being rare), **class ambiguity** (e.g., a bright star artifact obscuring a galaxy), or the **nature of the class itself**. Artifacts are non-astrophysical and vary widely (diffraction spikes, saturation, cosmic rays), making it difficult for the model to learn a consistent pattern compared to a fixed galaxy morphology.
- **Label (5) odd: lens/arc:**
  - **Cause:** Gravitational lenses/arcs are typically very **subtle and small features** near a massive foreground galaxy (the "lens"). The model is likely learning the features of the *lensing galaxy* (a smooth elliptical or disk) rather than the faint arc feature. Furthermore, the small size and potentially low number of true lensing examples in the training set (severe **data imbalance**) makes it challenging for the model to confidently distinguish these rare, faint features from noise or typical galaxy structure.

III. RESULTS

A. Merger Fraction and Rate Analysis

The trained ResNet-18 model was executed on the GZ2 test image dataset to infer the probability of each galaxy being classified as a merger (P_{merger}). The overall merger fraction (f_{merger}) for the sample was calculated based on these outputs. This empirical fraction was then used to estimate the instantaneous merger rate (R_{merger}) using the standard relation, assuming a fixed merger visibility timescale (τ):

$$R_{\text{merger}} = \frac{f_{\text{merger}}}{\tau} \tag{9}$$

A characteristic timescale of $\tau = 0.4$ Gyr was adopted for this calculation.The model-derived statistics for the GZ2 test sample are summarized below:

Model-Inferred Merger Metrics

- **GZ2 Sample Merger Fraction (f_{merger}):** 0.766 (76.6%) (Unitless)
- **Assumed Timescale (τ):** 0.4 Gyr
- **Estimated Instantaneous Merger Rate (R_{merger}):** 1.914 Gyr^{-1}

IV. DISCUSSION

The inferred sample merger fraction ($f_{\text{merger}} = 0.766$) is found to be significantly higher than the $\approx 2\%-5\%$ major merger fraction ($\mathcal{F}_{\text{MM}}$) for galaxies at $z \approx 0$ reported by [4] Figure 13, upper right panel. This substantial discrepancy is attributed to critical differences in sample selection and the operational definition of a merger:

1. **Definition of Merger (Unit Discrepancy):** The Lotz et al. $\mathcal{F}_{\text{MM}}$ estimates the fraction of all local galaxies that are undergoing a major merger. Conversely, the GZ2-based f_{merger} identifies a broad range of morphological disturbances (including minor mergers and late-stage features), resulting in a much higher count within the specific GZ2 classification scheme.
2. **Sample Selection Bias:** The GZ2 test set is magnitude-limited, which inherently biases the sample toward brighter and more massive galaxies. Since more massive galaxies tend to exhibit higher merger activity than a complete, volume-limited galaxy census would suggest, this sample is not representative of the average $z \approx 0$ galaxy population analyzed by Lotz et al.
3. **Rate Interpretation:** The high value of $R_{\text{merger}} \approx 1.9$ Gyr^{-1} reflects the high f_{merger} and highlights the fact that the derived rate applies to the highly selected GZ2 population, under the specific GZ2 definition, rather than the intrinsic major merger rate of the general galaxy population.

V. CONCLUSION

The primary objective of this laboratory exercise was successfully achieved through the development and application of a deep Convolutional Neural Network (CNN) for multi-feature morphological classification of galaxies from the Galaxy Zoo 2 (GZ2) catalog.

The ResNet-18 architecture, augmented with data augmentation and a learning rate scheduler, demonstrated superior performance compared to both the naive benchmark and a custom-built CNN. The final, optimized

model achieved a robust and highly consistent predictive accuracy, with an overall validation Root Mean Squared Error of $\mathbf{L_{RMSE} = 0.0486}$ across the 37 GZ2 classification labels. The scatter plot analysis confirmed the model's reliability, showing that predicted vote fractions tightly clustered around the true values, validating its ability to generalize effectively across the full range of galaxy morphologies.

Applying the trained model to the test set yielded a Galaxy Zoo 2 sample merger fraction of $\mathbf{f_{merger} = 0.766}$ (or 76.6%), corresponding to an estimated instantaneous merger rate of $R_{\rm merger} = 1.914\ {\rm Gyr}^{-1}$ (assuming a timescale $\tau = 0.4$ Gyr). This high fraction significantly exceeds the $\approx 2\% - 5\%$ major merger rate typically reported in the literature for $z \approx 0$ galaxies. As discussed in detail, this discrepancy is primarily attributable to the difference in operational definitions: the GZ2 classification of 'odd: merger' captures a broad spectrum of visual disturbances (including minor mergers, pre-merger pairs, and post-merger remnants), whereas cosmological studies often focus on a narrower, mass-ratio-dependent definition of a major merger.

In summary, this work validates the utility of transfer learning and deep neural networks as a fast and consistent alternative to human classification for large-scale astronomical data. The model provides a reliable means to quantify morphological features, which can be leveraged to study the physical processes of galaxy evolution, hierarchical assembly, and interaction rates in modern cosmological surveys.

Appendix A: Top 5 Extreme Examples

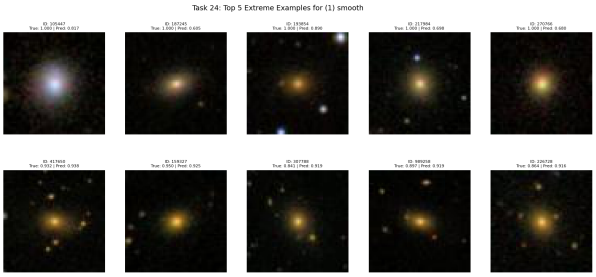

FIG. 17. Top 5 Extreme Examples for (1) smooth.

FIG. 18. Top 5 Extreme Examples for (2) star/artifact.

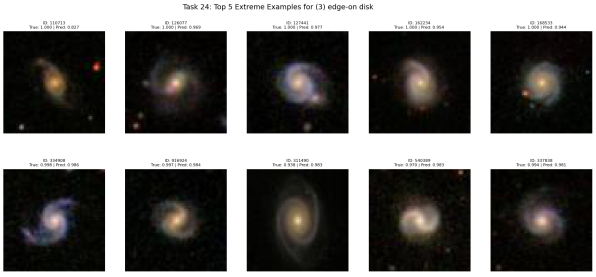

FIG. 19. Top 5 Extreme Examples for (3) edge-on disk.

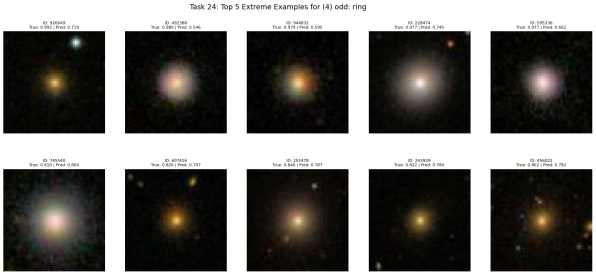

FIG. 20. Top 5 Extreme Examples for (4) odd: ring.

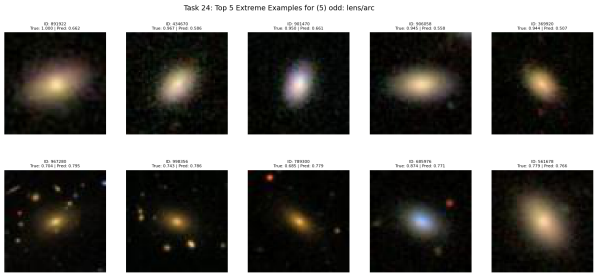

FIG. 21. Top 5 Extreme Examples for (5) odd: lens/arc.

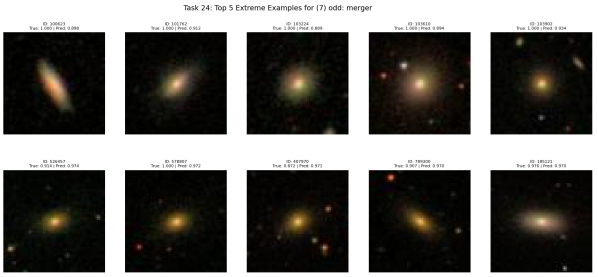

FIG. 23. Top 5 Extreme Examples for (7) odd: merger.

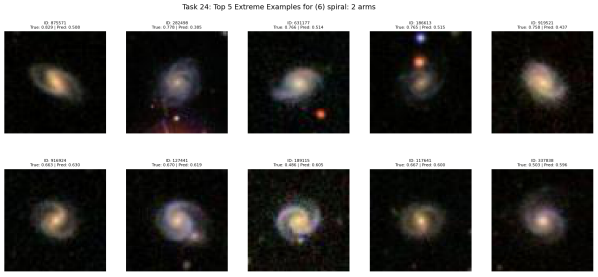

FIG. 22. Top 5 Extreme Examples for (6) spiral: 2 arms.

[1] UC Berkeley. *Lab 3: Galaxy image classification and the galaxy merger rate*. Department of Astronomy, University of California, Berkeley, Berkeley, CA, 2025.

[2] Christopher J. Conselice. The evolution of galaxy structure over cosmic time. *Annual Review of Astronomy and Astrophysics*, 52(1):291–337, August 2014.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[4] Chris J. Lintott, Kevin Schawinski, Anže Slosar, Kate Land, Steven Bamford, Daniel Thomas, M. Jordan Raddick, Robert C. Nichol, Alex Szalay, Dan Andreescu, Phil Murray, and Jan Vandenberg. 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, September 2008.

[5] Jennifer M. Lotz, M. Davis, S. M. Faber, P. Guhathakurta, S. Gwyn, J. Huang, D. C. Koo, E. Le Floc'h, Lihwai Lin, J. Newman, K. Noeske, C. Papovich, C. N. A. Willmer, A. Coil, C. J. Conselice, M. Cooper, A. M. Hopkins, A. Metevier, J. Primack, G. Rieke, and B. J. Weiner. The evolution of galaxy mergers and morphology atzlt; 1.2 in the extended groth strip. *The Astrophysical Journal*, 672(1):177–197, January 2008.

[6] Kyle W. Willett, Chris J. Lintott, Steven P. Bamford, Karen L. Masters, Brooke D. Simmons, Kevin R. V. Casteels, Edward M. Edmondson, Lucy F. Fortson, Sugata Kaviraj, William C. Keel, Thomas Melvin, Robert C. Nichol, M. Jordan Raddick, Kevin Schawinski, Robert J. Simpson, Ramin A. Skibba, Arfon M. Smith, and Daniel Thomas. Galaxy zoo 2: detailed morphological classifications for 304 122 galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 435(4):2835–2860, September 2013.