

Galaxy Image Classification using CNN and ResNet-50 with Stratified K-Fold

Dataset Description:

The **Galaxy10 DECaLS** dataset contains 17,736 colored galaxy images of shape **256×256×3** in 10 classes. These images are from the DESI Legacy Imaging Surveys, with class labels obtained from Galaxy Zoo.

Only the images were used in this project. The other features (ra, dec, redshift, pxscale) were ignored.

Class Distribution (Original):

- Class 0 (Disturbed Galaxies): 944
- Class 1 (Merging Galaxies): 1616
- Class 2 (Round Smooth Galaxies): 2165
- Class 3 (In-between Round Smooth Galaxies): 1840
- Class 4 (Cigar Shaped Smooth Galaxies): 290
- Class 5 (Barred Spiral Galaxies): 1860
- Class 6 (Unbarred Tight Spiral Galaxies): 1551
- Class 7 (Unbarred Loose Spiral Galaxies): 2366
- Class 8 (Edge-on Galaxies without Bulge): 1264
- Class 9 (Edge-on Galaxies with Bulge): 1553

The class imbalance was handled by augmenting underrepresented classes, bringing every class to 2,366 samples.

Preprocessing Summary:

- Removed nulls and outliers using **z-score thresholding** ($z > 2$).
- Shuffled the dataset.
- Maintained original image color channels (no grayscale conversion).
- Applied **data augmentation**.

Model 1: Convolutional Neural Network (CNN):

Architecture:

Conv2D (32) → MaxPool

Conv2D (64) → MaxPool

Conv2D (128) → MaxPool

Flatten → Dense (128) → Dropout (0.5) → Dense (10)

- Optimizer: Adam
- Loss: Sparse Categorical Cross entropy
- Callback: Early Stopping (patience=5)

Performance (on test set):

- **Accuracy:** 14%
- **Observation:** Only Class 2 was predicted (recall = 1.0), others had 0 precision/recall.
- The CNN failed to generalize and showed severe overfitting.

♦ Classification Report:				
	precision	recall	f1-score	support
0	0.0000	0.0000	0.0000	189
1	0.0000	0.0000	0.0000	323
2	0.1401	1.0000	0.2458	433
3	0.0000	0.0000	0.0000	368
4	0.0000	0.0000	0.0000	58
5	0.0000	0.0000	0.0000	372
6	0.0000	0.0000	0.0000	310
7	0.0000	0.0000	0.0000	473
8	0.0000	0.0000	0.0000	253
9	0.0000	0.0000	0.0000	311
accuracy			0.1401	3090
macro avg	0.0140	0.1000	0.0246	3090
weighted avg	0.0196	0.1401	0.0344	3090

Figure 1 Classification report for CNN

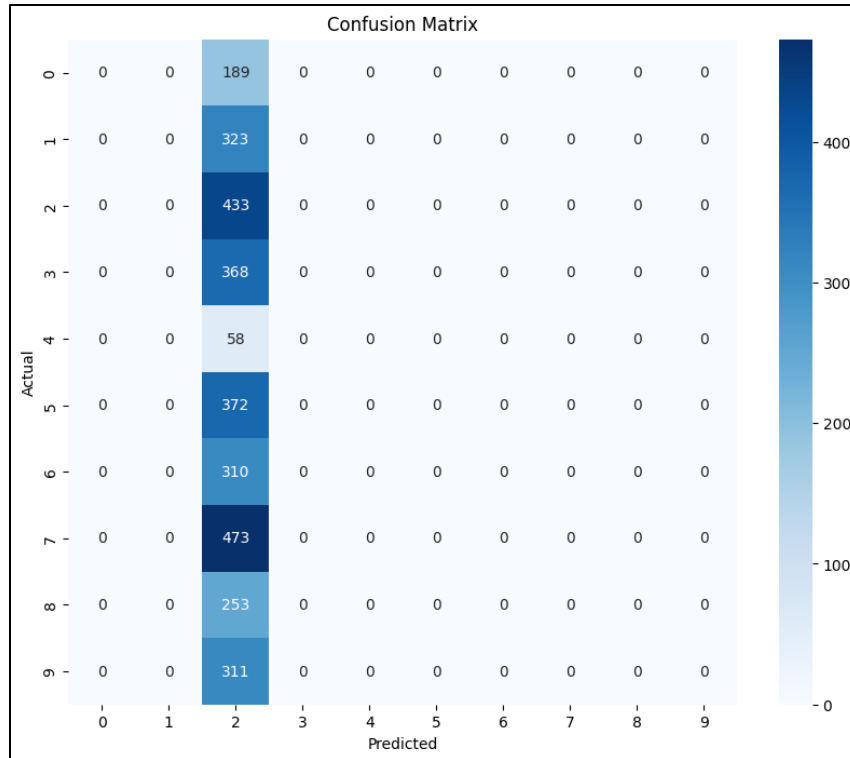


Figure 2 Confusion Matrix for CNN

Model 2: ResNet-50 with Stratified K-Fold:

Used **PyTorch** + **pretrained ResNet-50**. Modified final FC layer to match 10 classes.

Training Strategy:

- **Stratified K-Fold ($k = 4$)** for balanced class splits.
- Used only augmented & balanced data.
- **Input resize:** 224×224
- **Epochs:** 10 per fold
- **Batch Size:** 32
- Optimizer: Adam ($lr = 1e-4$)
- Loss: CrossEntropyLoss

For each fold:

- Model was trained and saved as .pt file.
- Performance was evaluated using classification report, confusion matrix, and plots.

Model Performance:

- All folds performed similarly; **Fold 4** performed slightly better and was selected.
- Accuracy on unseen test set of 1500 samples: **84%**

Classification Report:					
	precision	recall	f1-score	support	
0	0.7518	0.6770	0.7125	452	
1	0.8123	0.9615	0.8806	441	
2	0.9142	0.9375	0.9257	432	
3	0.9235	0.8578	0.8894	436	
4	0.9344	0.9124	0.9232	468	
5	0.8910	0.8421	0.8659	437	
6	0.8196	0.8326	0.8260	442	
7	0.7482	0.7324	0.7402	426	
8	0.9180	0.9159	0.9169	452	
9	0.9124	0.9574	0.9344	446	
accuracy			0.8630	4432	
macro avg	0.8625	0.8627	0.8615	4432	
weighted avg	0.8631	0.8630	0.8619	4432	

Figure 3 Classification Report for resnet-50 (k=4)

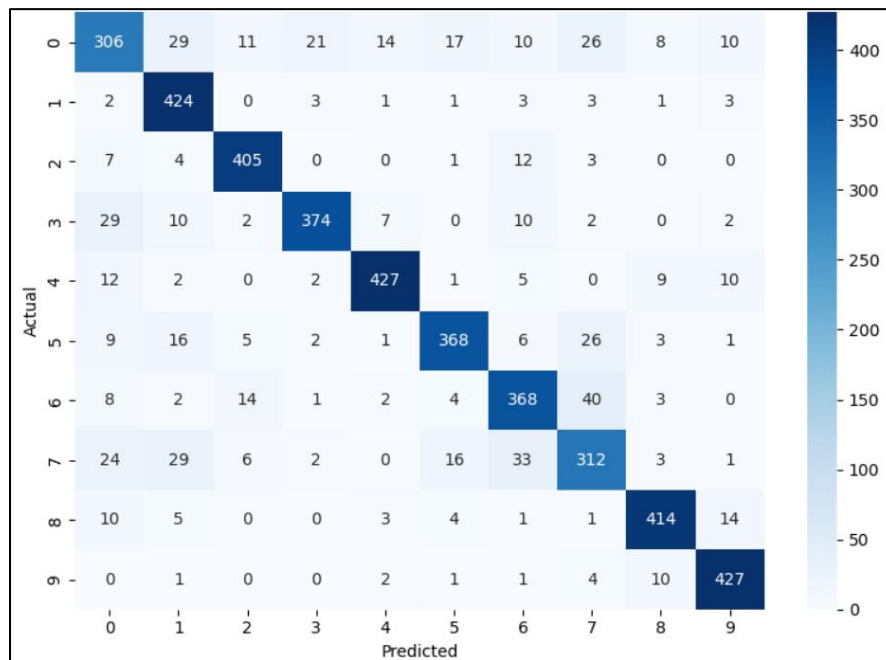


Figure 4 Confusion Matrix for Resnet-50 (k=4)

Evaluated on 1500 unseen test samples. The histogram below shows the final class-wise distribution of predictions.

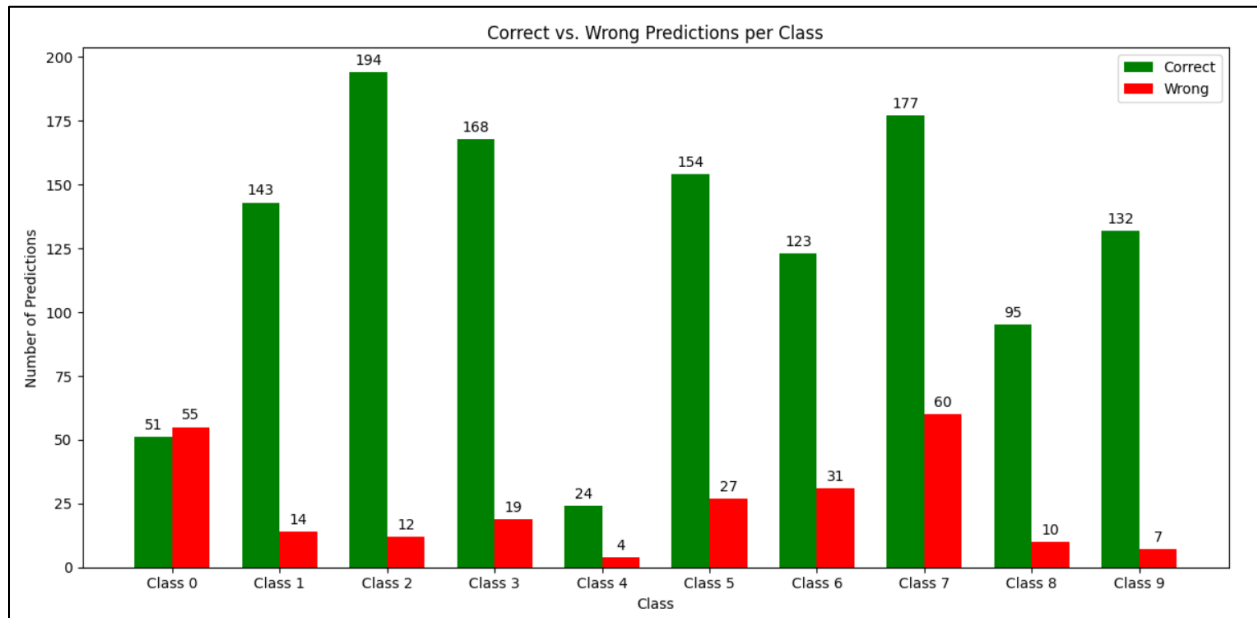


Figure 5 Predictions on unseen data

Conclusion:

- Basic CNN was insufficient for the complexity of the galaxy image classification task.
- ResNet-50, with pretrained weights and Stratified K-Fold training, achieved **robust generalization**.
- Final accuracy: **~84%** on unseen samples.

This model can now be used for automated galaxy morphology classification in astronomy-related workflows.