A Machine-Learning Approach to Galaxy Morphology and Parameter Estimation

Yash R. Bhora, Benjamin Ecsedy

Introduction

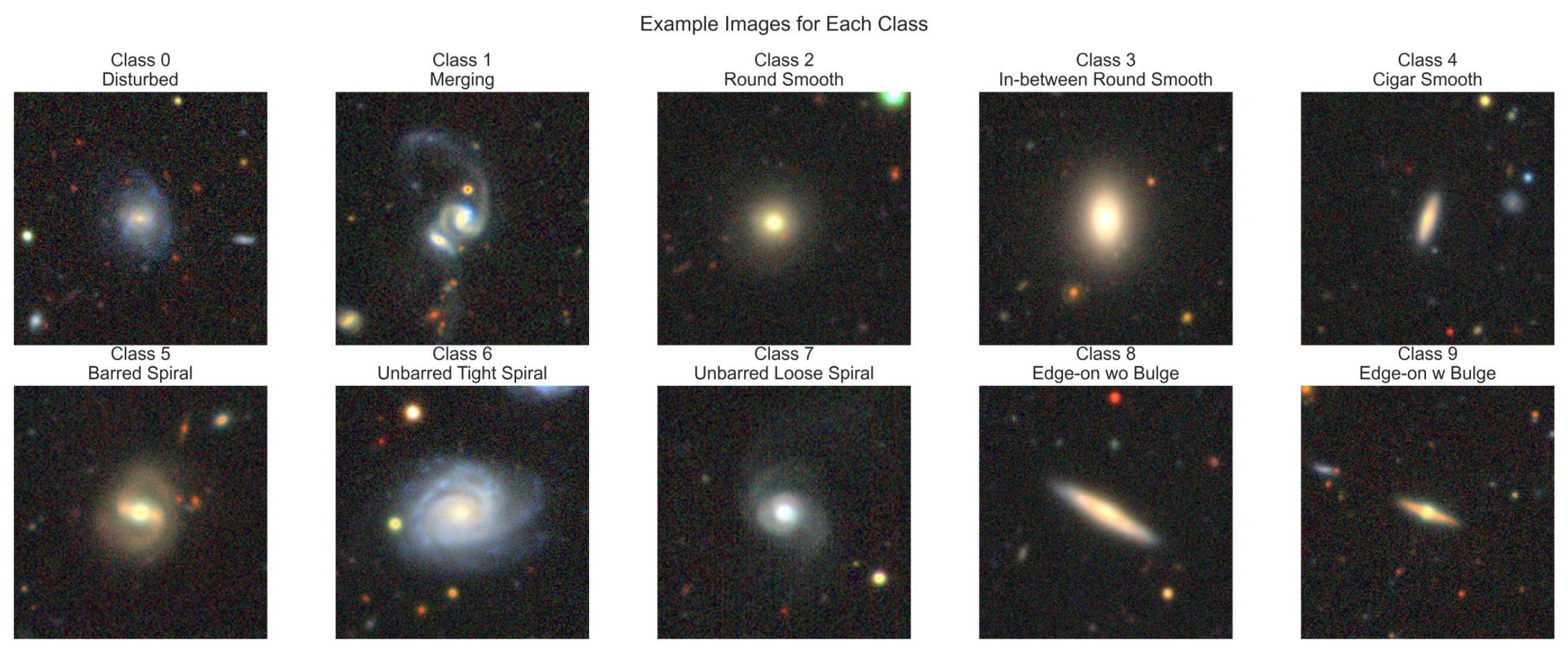
- Galaxy morphology (e.g., spiral, elliptical, or irregular) and classification are essential for understanding the structure and evolution of galaxies in the universe.
- Accurately estimating a galaxy's mass is key to tracking the growth and distribution of matter across the universe.
- Traditional analysis methods rely heavily on manual classification, while machine learning offers scalable solutions by learning complex patterns from multi-parameter datasets of galaxy measurements.
- Machine learning enables astronomers to efficiently analyze massive datasets, helping uncover deeper cosmological insights with high accuracy.
- Hubble's Law states that distant galaxies are moving away from Earth due to the expansion of the universe, causing redshift in the light that is emitted by galaxies, in line with the Doppler Effect.
- Photometric redshift estimation is a key technique to estimate the redshift of a galaxy but has limitations in accuracy and reliability using traditional methods, so other methods need to be employed.

Data & Goals

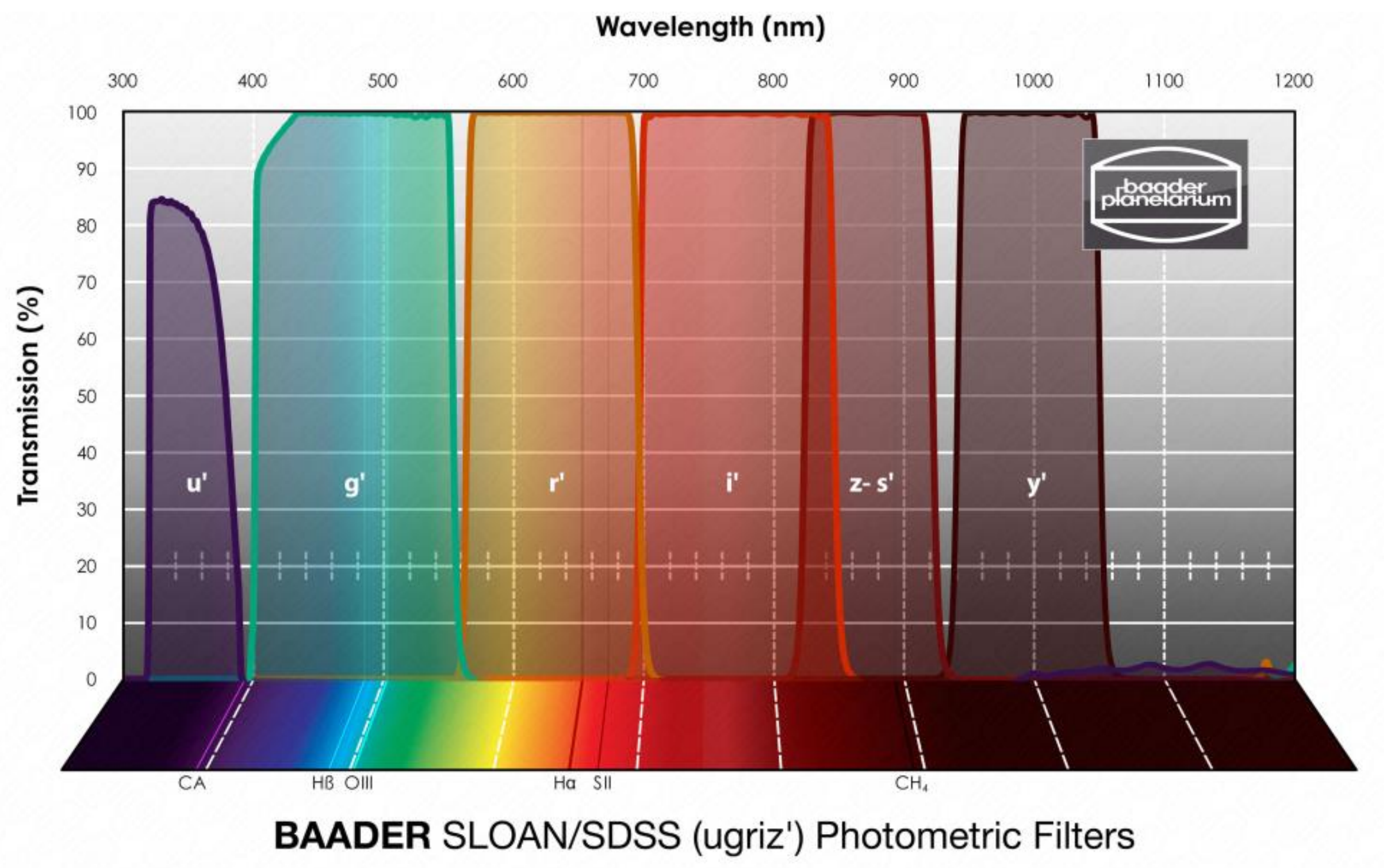
Primary Goal: Evaluate the impact of data augmentation and model complexity on classification and prediction.
The data was obtained using the *Galaxy DECaLS* Dataset (Leung & Bovy, 2019), containing *17736 256x256* pixels colored galaxy images (g, r and z band) for 10 morphologies shown below, as well as a set of simulated data from the Rubin Observatory of galaxy magnitude measurements of different color filters in six astronomical bands using the Buzzard V-1.0 simulation.

The main tasks at hand were to:

1. Compare performance using "raw" images vs enhanced with data augmentation.
2. Assess performance gains among three CNN architectures: SimpleCNN, PowerfulCNN, and EfficientNet_B2 (transfer-learning)
3. Determine the predictive power of a simplified statistical model on astronomical data.



Example images from each morphological class, showcasing the diverse appearances of galaxies used for training.



Wavelengths spectra of light captured in each of the six color filters (u, g, r, I, z, and y).

Methodology

Morphology Classification with CNNs and Data Augmentation

Data Loading & Preprocessing: Read images from HDF5 file. We then created training, validation and test sets using a 70/15/15.

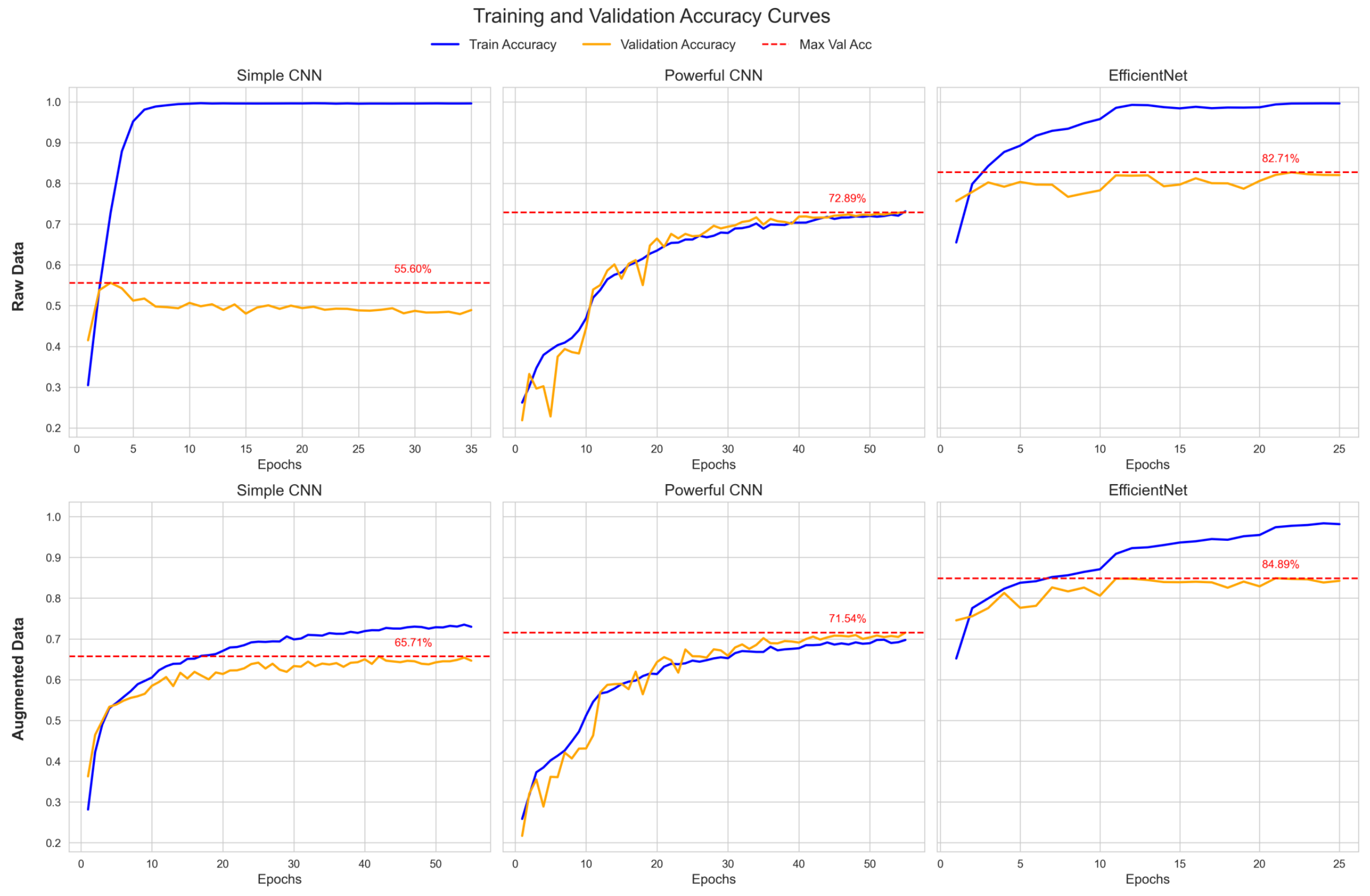
Transformations: Both raw and augmented images were normalized according to ImageNet standards. For augmentations, random horizontal flips, rotations (15 degrees), and random resized crops (scale 0.8-1.0) were applied prior to normalization.

Model Implementation and Training: We utilized PyTorch for training, using an Adam optimizer with an initial learning rate of 0.001 and a scheduler that decays the learning rate (by 0.5 every 10 epochs). Depending on convergence, we trained the models for up to 55 epochs. The best model was saved throughout training based on validation accuracy.

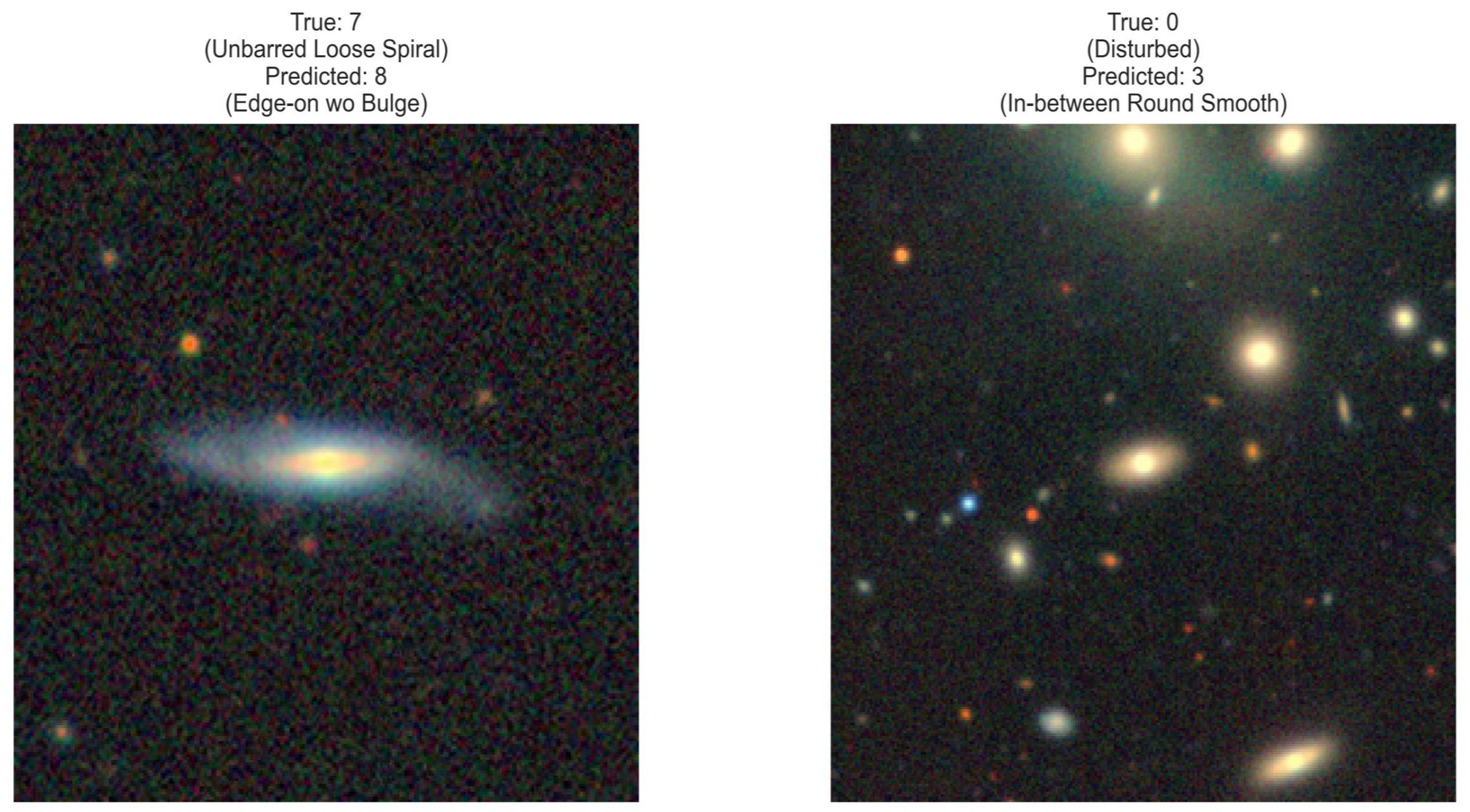
SimpleCNN: A lightweight network with two convolutional layers followed by max pooling and two fully connected layers. The model was trained from scratch (no pretraining) to serve as a baseline classification.

PowerfulCNN: A deeper, more robust network that includes multiple convolutional blocks with batch normalization, ReLU activations, and max pooling. Incorporates an adaptive average pooling layer to reduce the spatial dimensions before the classifier, making it more resilient to variations in input image resolution. Also trained from scratch.

EfficientNet_B2: Utilizes a pre-trained model to leverage features learned on large-scale datasets for transfer learning. Fine-tuning adapts the high-level representations to the galaxy classification task, generally providing improved performance with faster convergence.



Accuracy curves comparing raw vs. augmented data across the three CNN models over training epochs.



Examples of misclassifications in the galaxy image classifier

Mass and Redshift Estimation with Bayesian Modeling

Data Composition: The data consists of simulated observations of 111,172 galaxies under six different optical filters: *u*, *g*, *r*, *i*, *z*, and *y*, the error in each of these measurements, the galaxy mass, and the galaxy redshift.

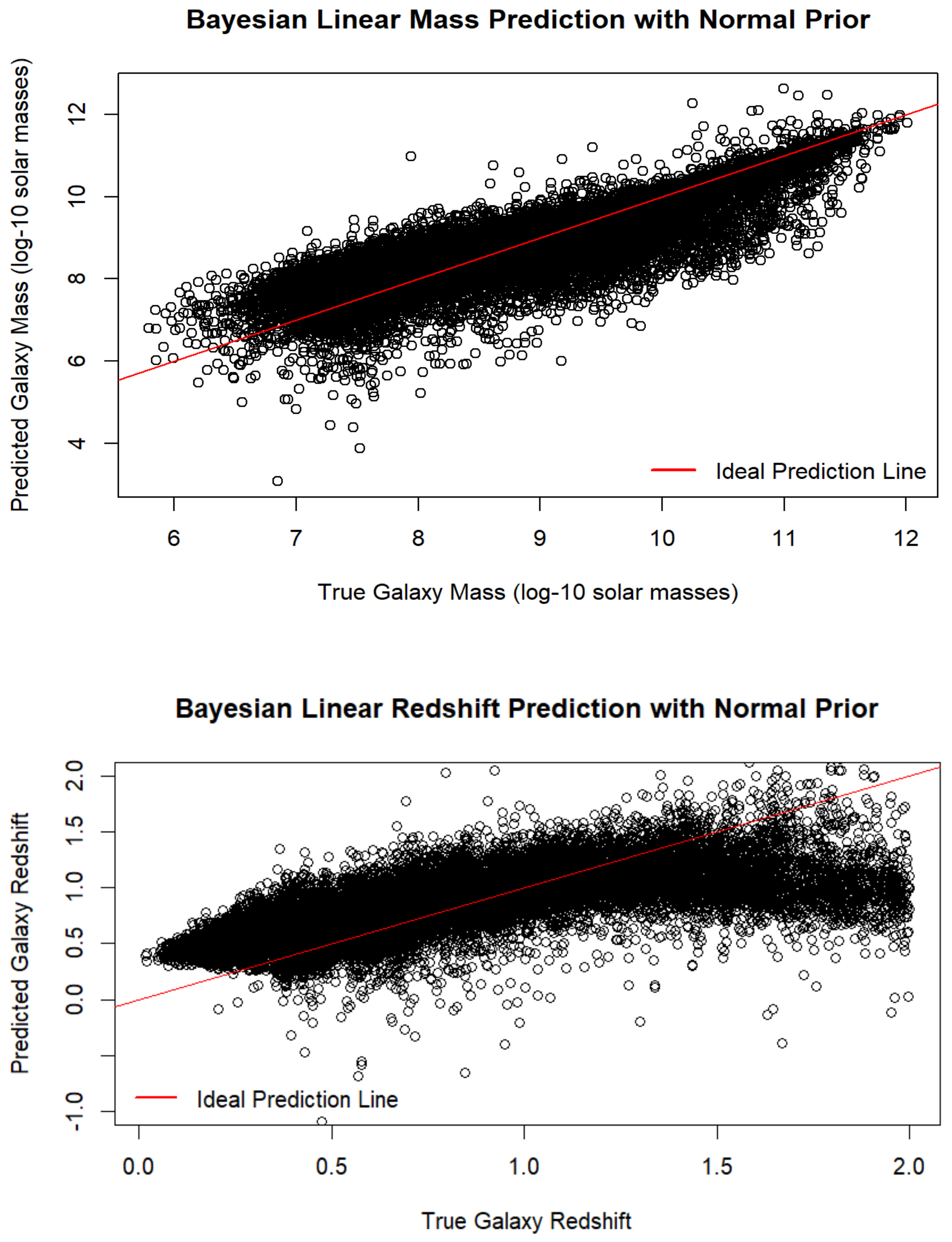
Data Loading & Preprocessing: The color filter data was split into an 80/20 train/test split and the Bayesian regressor was trained on the train set, with the predictions on the test set shown below.

Prior Distribution: Benítez (2000) specifies an appropriate prior distribution to use for estimating the redshift of a galaxy from a given set of color magnitudes m_0, given by $p(z|m_o) \propto z^\alpha \exp((-z/(z_0 + k(m_0-20)))^\alpha)$, where α, k, and z_0 are shape parameters. By estimating a prior distribution, data from each galaxy can be used to update the predictions of the distribution of galaxy mass and redshift given observation data to influence predictions for other galaxies with unknown masses and redshifts.

Prior Estimation: This investigation aimed to determine whether a simpler model for the prior distribution would still achieve accurate results. Instead of implementing the Benítez prior, a normal distribution of each magnitude was assumed.

Feature Engineering: To construct the Bayesian predictor, we wanted to capture both linear and nonlinear dependence of the magnitude values on the mass and redshift of a galaxy. Thus, each of the mass and redshift was constructed to be determined by the features using a polynomial of degree 2, taking as inputs the magnitude from each color filter, the degree-2 terms of these magnitudes (e.g. *gz*, *ui*, y^2), and a constant bias term.

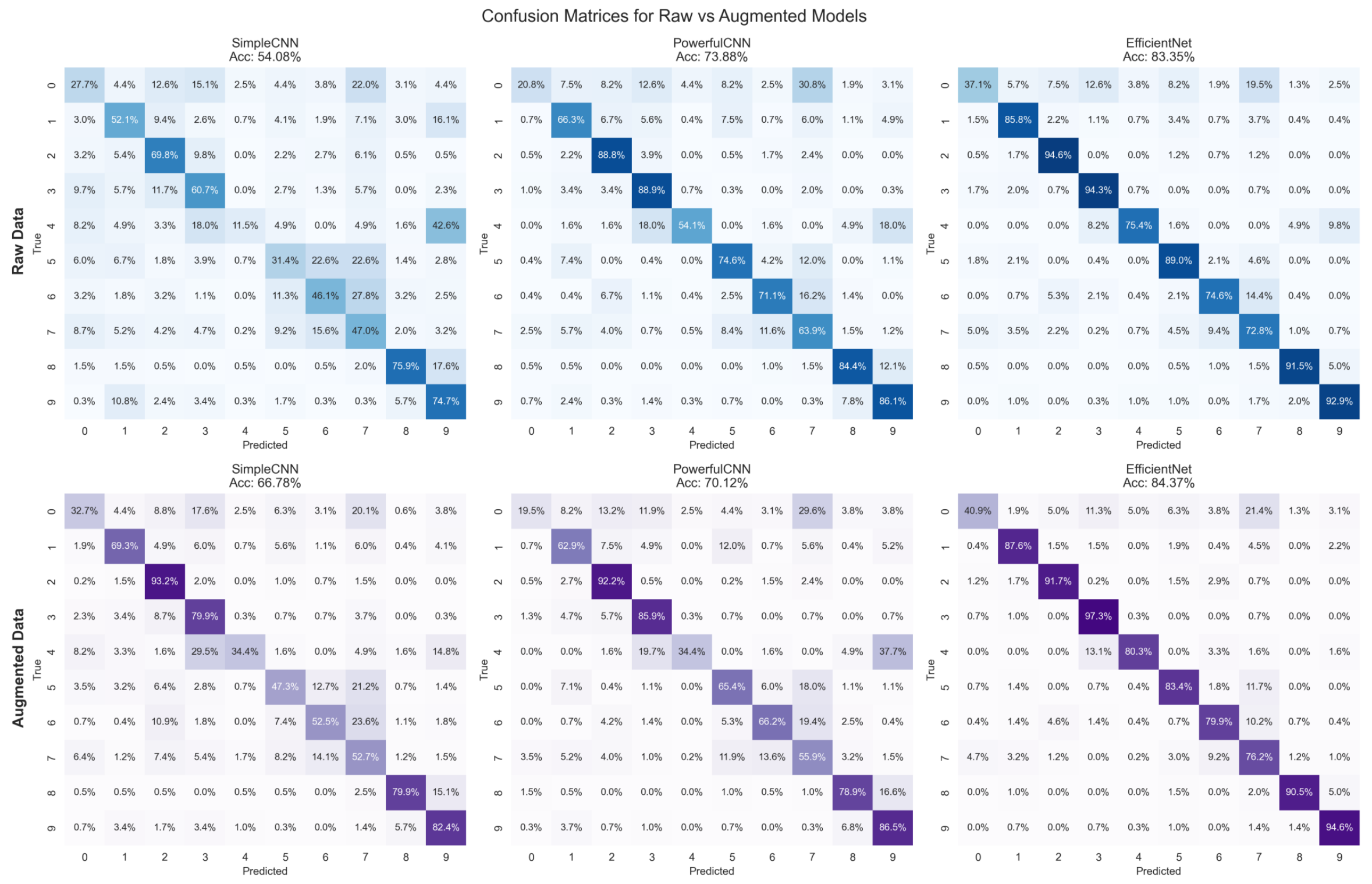
Posterior Estimation: The posterior distribution of the regression coefficients is computed analytically using matrix operations, resulting in both a mean prediction and an uncertainty. These means and uncertainties were compared with the true value of the galaxy mass and redshift to assess the performance of the Bayesian model.



True versus predicted mass and redshift for each value in the test dataset.

Results & Conclusion

Galaxy Classification Results

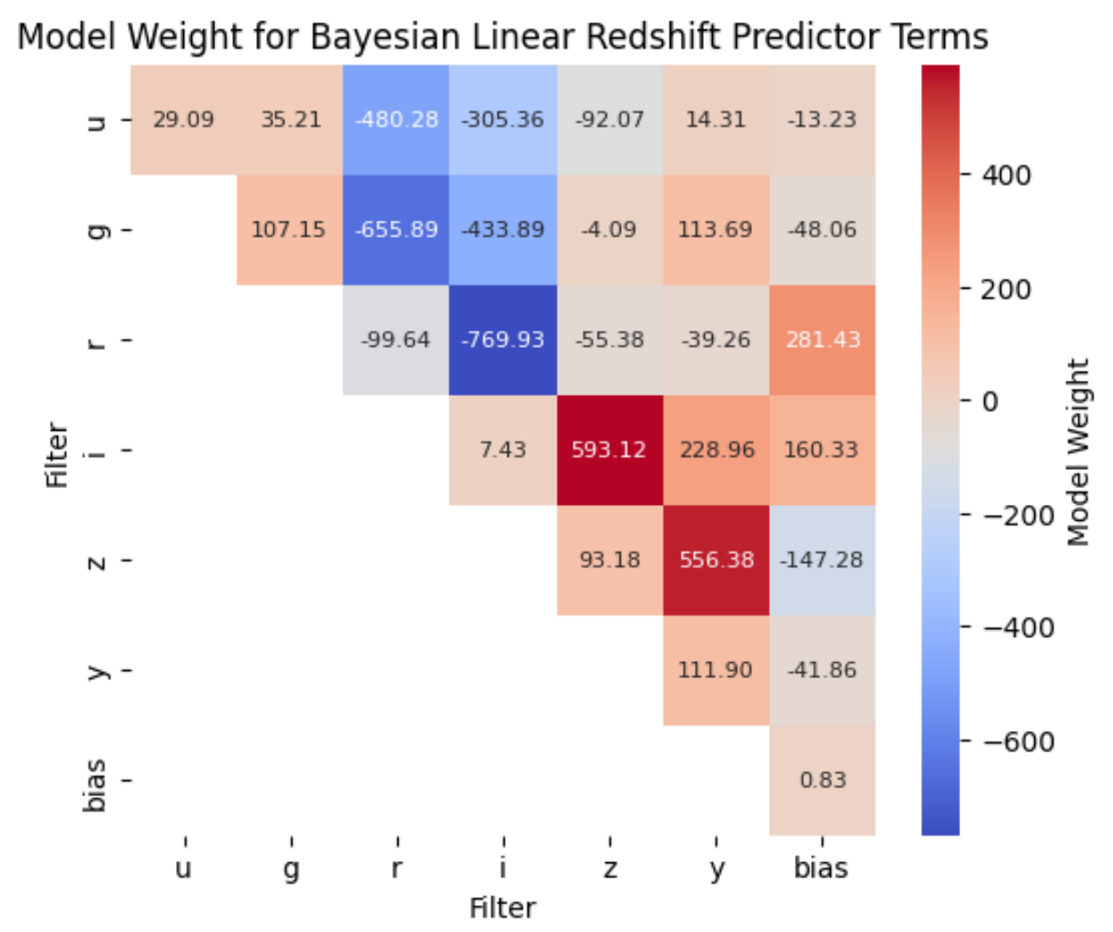


Confusion matrices comparing raw vs. augmented training across three CNN models.

- **Augmentation Boosts Performance:** Models trained with augmented data always yielded higher overall accuracy to their 'raw' counterparts.
- **EfficientNet Converges Quickly**: As a transfer learning model, EfficientNet reaches peak validation accuracy in significantly fewer epochs than the other models.
- **Overfitting Resistance**: SimpleCNN overfits to raw data quickly. PowerfulCNN avoids overfitting (training and validation curves are close), even better than EfficientNet.
- **Improved Cigar Galaxy Classification**: Unlike the other models, EfficientNet not only performs the best, but accurately classifies "Cigar Smooth" galaxies despite its low fraction in the dataset.

Mass and Redshift Prediction Results

- **Model Performance:** The mean squared error between the true and predicted galaxy mass was 0.304, while that of the true and predicted redshift was 0.0895, both when evaluated on the test set. These low MSEs show that the model does a good job at predicting the relationship between the true and predicted values.
- **Feature Weights:** The feature weight were all very large compared to the bias term, indicating a potential of overfitting to the training data, although the model did perform well on the test set.
- **Bias Interpretation**: The bias weight mean, 0.83, represents a "baseline" estimation measurement for the redshift of a galaxy, without any knowledge of its color magnitude.
- **Benítez Prior and Sampling:** Since the Gaussian distribution was only an approximation for the Benítez prior, a better approximation or the Benítez prior itself could be used to improve the predictor results.



Mean parameter weights for each cross-term

References

Agena Astro. Baader SLOAN/SDSS (ugriz') Photometric Filter Set - 1.25" Mounted # FSLNSET-1 2961700.
Benítez, N. (1998). Bayesian photometric redshift estimation. The Astrophysical Journal, (Vol. 536, Issue 2, pp. 571-583).
Dieleman, S., Willett, K. W., & Dambre, J. (2015). Rotation-invariant convolutional neural networks for galaxy morphology prediction. Monthly Notices of the Royal Astronomical Society, 450(2), 1441–1459.
Leung, H. W., & Bovy, J. (2019). Galaxy10 DECaLS: A CIFAR10-like dataset for galaxy morphology classification [Data set]. Zenodo.
Schmidt, S. J. (2020). Evaluation of probabilistic photometric redshift estimation approaches for The Rubin Observatory Legacy Survey of Space and Time (LSST). Monthly Notices of the Royal Astronomical Society, (Volume 499, Issue 2, pp. 1587–1606).
Schneider, J., Stenning, D. C., & Elliott, L. T. (2023). Efficient galaxy classification through pretraining. Frontiers in Astronomy and Space Sciences, 10.
Sloan Digital Sky Survey, What is Color?, SkyServer DR1.
Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning (Vol. 97, pp. 6105–6114). PMLR.