A Machine-Learning Approach to Galaxy Morphology and Parameter Estimation

Yash R. Bhora, Benjamin Ecsedy

Introduction and Literature Review

Understanding the diversity of galaxies, in both their visible shapes and their physical properties, is a central challenge in astronomy. Galaxy morphology classification (identifying galaxies as spirals, ellipticals, irregulars, etc.) is essential for studying how galaxies form and evolve. Likewise, accurately estimating a galaxy's stellar mass and redshift (a proxy for distance/velocity) is crucial for tracing the growth of cosmic structures and the expansion of the universe.

Traditionally, astronomers classified galaxy images by eye and measured distances via spectroscopy, but manual approaches are time-consuming and cannot keep pace with modern survey data. Projects like *Galaxy Zoo* enlisted volunteers to classify over a million galaxies, yielding valuable datasets; however, even crowdsourcing does not scale well enough to handle the ever-growing flood of images from current and upcoming telescopes (Dieleman et al., 2015). Similarly, photometric redshift estimation (inferring redshift from multi-band brightness instead of spectra) has historically suffered from limited accuracy and reliability with classical methods. There is a pressing need for automated techniques that can learn from data and efficiently analyze these large astronomical datasets.

Machine learning has emerged as a powerful solution to these scalability challenges. In particular, convolutional neural networks (CNNs), specialized deep learning models for images, can learn complex visual patterns and have demonstrated human-level performance in galaxy morphology classification. At the same time, ML-based regression methods (including probabilistic approaches like Bayesian regression) offer a way to predict continuous galaxy parameters (e.g. redshift, mass) from photometric measurements, often more quickly or flexibly than traditional fitting techniques.

In this project, we explore a two-pronged ML approach to galaxy analysis. On one hand, we apply CNN-based models to galaxy images from the Galaxy10 DECaLS dataset to classify galaxy morphology. We compare a simple custom CNN, a more complex/deeper CNN, and a state-of-the-art pre-trained network to evaluate how model complexity and transfer learning affect classification performance. On the other hand, we use a Bayesian regression model to estimate galaxy parameters, specifically stellar mass and photometric redshift, from simulated multi-band observations. This Bayesian model incorporates prior information and yields uncertainty estimates, which is advantageous for photometric redshift prediction where probabilistic approaches have been shown to improve reliability.

In the past decade, CNNs have revolutionized how astronomers classify galaxy morphologies. Early efforts to automate morphology (pre-dating deep learning) struggled to reach the accuracy of human classifiers. The advent of projects like Galaxy Zoo, which provided millions of volunteer-labeled galaxy images, enabled supervised training of deep CNNs to recognize galaxy features. Notably, Dieleman et al. (2015) trained a rotation invariant CNN on Galaxy Zoo data and achieved near-human performance, reproducing volunteer consensus with >99% accuracy on images where the galaxy's morphology was clear. As datasets grew, researchers have expanded classification schemes to finer morphology classes (beyond simple spiral/elliptical) using deeper networks. Overall, the literature demonstrates that CNN-based classifiers can handle the scale and complexity of modern surveys, turning what was once a manual, bottleneck process into a largely automated one without sacrificing accuracy.

Two technical practices, transfer learning and data augmentation, have been key to improving CNN performance on relatively limited astronomical image datasets. *Transfer learning* involves taking a neural network pre-trained on a large generic image database (like ImageNet) and fine-tuning it on astronomy images. This approach leverages

learned low-level features (edges, shapes, textures) from millions of everyday images and applies them to galaxy imagery. Recent studies have widely adopted transfer learning for galaxy morphology classification (Schneider et al., 2023). For instance, Domínguez Sánchez et al. (2018) demonstrated that using a pre-trained deep CNN (originally trained on natural images) significantly improved accuracy in classifying Sloan Digital Sky Survey galaxy images, while also reducing the training time required. In this project, the EfficientNet model (pre-trained on ImageNet) exemplifies this trend of bringing in external learned knowledge to boost astronomical classification. Data augmentation is another widely used technique: by applying random transformations to images (rotations, flips, slight rescaling or noise), one can massively increase the effective training sample and make the model invariant to those transformations. In galaxy morphology tasks, a common practice since Dieleman et al. (2015) is to generate many rotated versions of each galaxy image to account for random orientation. In our work, we incorporate data augmentation during CNN training to enhance performance, especially for the simpler CNN models that might otherwise struggle with limited data.

As for previous work on redshift estimation, Narciso Benítez's 1998 paper "Bayesian photometric redshift estimation" provides multiple Bayesian modeling architectures for predicting a galaxy's redshift given its measured magnitude under different filters. In it, he details a statistical method for applying Bayesian probability to photometric redshift estimation, BPZ, which explicitly uses prior probability distributions when predicting galactic redshift, which is often ignored by other methods. Benítez thus frames the problem of finding a galaxy's redshift not as one of finding its value, but rather the distribution of potential redshifts z, conditional on the input vector m_0 of different magnitude (brightness) measurements in certain color filters. The prior specified in the paper is given as $p(z|m_0) \propto z^{\alpha} exp((\frac{-z}{z_0+k(m_0-20)})^{\alpha})$, where α, k, and z_0 are shape parameters. The shape parameters differ slightly for each type of galaxy, but the derivations in the Benítez paper help make this model as generalizable as possible to any arbitrary galaxy.

The work of Jasche and Wandelt (2012) expanded on this work, demonstrating the feasibility and accuracy of applying an isotropic boundary condition on the deviation of the parameters involved in the Bayesian analysis of redshift, showing that accurate results could be obtained assuming an isotropic distribution even if there are high uncertainties in the given measurements. Thus, the combination of the Bayesian methodology outlined in Benítez and the accuracy of simplified Bayesian models shown in Jasche and Wandelt inspired the investigation into whether even simpler Bayesian models can accurately predict the redshift characteristics of galaxies given input magnitude data.

In summary, previous research provides a strong foundation for our project. CNN-based models, especially when enhanced by transfer learning and ample data augmentation, have proven highly effective for galaxy morphology classification, even as the data volume grows into the millions. Likewise, machine learning and Bayesian techniques have shown great promise in predicting galaxies' numerical parameters from photometry, often outperforming classical methods and scaling to the needs of next-generation surveys. By drawing on these advancements, our project implements and compares state-of-the-art approaches in a unified study. The goal is to demonstrate how modern machine learning can handle both image-based classification and multi-variable regression in astrophysics and to evaluate the benefits of network choice, transfer learning, and Bayesian modeling in the context of a class research project. The following sections detail our methodology and data, and how the literature's guidance translated into our experimental design.

Machine Learning Methodology and Data Collection

Galaxy Morphology Classification with Convolutional Neural Networks

Data Selection and Preparation: For galaxy morphology classification, we used the *Galaxy10 DECaLS* dataset (Leung & Bovy, 2019). This dataset was chosen not only because of its public access but also because of its comprehensive coverage and improved image quality compared to its predecessor (Galaxy10), thus allowing us to more accurately distinguish among different galaxy types. The dataset combines image data from the DESI Legacy Imaging Surveys (DECaLS) with human-validated labels derived from volunteer classifications in the Galaxy Zoo project. This dataset consists of 17, 736 color galaxy images of dimension 256x256 pixels across three photometric bands (g, r, and z). The galaxies are divided into 10 distinct morphological classes, including categories such as "Disturbed," "Merging," "Round Smooth," various spiral formations, and two types of edge-on galaxies (with and without central bulges), as illustrated in Figure 1. We partitioned the dataset into training, validation, and test subsets following a 70/15/15 ratio, providing balanced sets for effective model training, hyperparameter tuning, and unbiased evaluation.

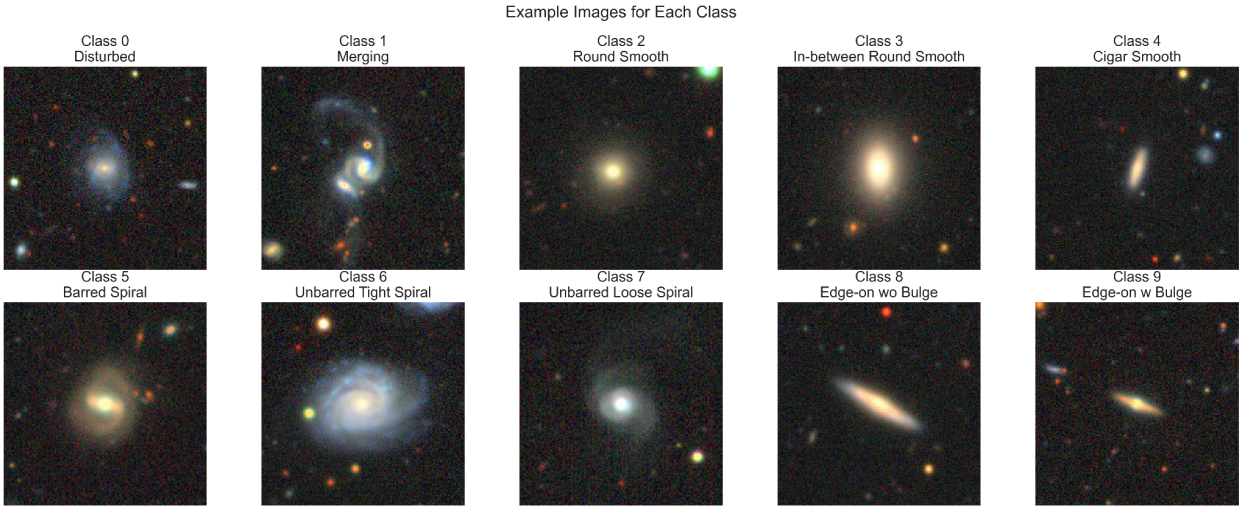

Figure 1: Example images illustrating the 10 distinct morphology classes from the Galaxy10 DECaLS dataset.

CNN Architectures Evaluated: To classify galaxies according to morphology, we explored three distinct Convolutional Neural Network (CNN) architectures: SimpleCNN, PowerfulCNN, and EfficientNet. The former two were trained from scratch whereas EfficientNet is a pre-trained model that was fine-tuned to the galaxy data.

SimpleCNN served as our baseline model, built within the scope of our course's CNN implementations. It comprises two convolutional layers, each followed by a ReLU activation and max pooling, progressively reducing spatial dimensions from 256x256 pixels to 64x64 pixels. The flattened feature maps are then processed by two fully connected layers, ultimately classifying galaxies into the 10 morphological categories. Due to its straightforward design, SimpleCNN provides insight into baseline performance with minimal computational complexity.

Extending beyond standard course implementations, we developed a more advanced CNN called PowerfulCNN, intended to capture more complex morphological patterns. PowerfulCNN features three convolutional blocks, each containing two convolutional layers accompanied by batch normalization and ReLU activation functions. Spatial dimensions progressively decrease through max pooling layers. An adaptive average pooling layer further condenses the feature maps to fixed dimensions, irrespective of input image size. This is followed by dropout layers and a

two-layer fully connected classifier, significantly reducing overfitting and improving the model's ability to distinguish subtle morphological differences.

Lastly, we employed a state-of-the-art approach via EfficientNet_B2, leveraging transfer learning techniques. EfficientNet uses compound scaling, efficiently balancing network depth, width, and resolution. It features Mobile Inverted Bottleneck Convolutional (MBConv) blocks, enriched with squeeze-and-excitation modules that dynamically recalibrate channel-wise feature responses (Tan & Le, 2019). Pretrained on the extensive ImageNet dataset, EfficientNet benefits significantly from these learned general-purpose representations, enabling rapid convergence and superior performance even with fewer training epochs compared to models trained from scratch.

Data Augmentation and Training: To further enhance performance and model generalizability, we experimented with data augmentation strategies. Raw data preprocessing involved only standard normalization (mean and standard deviation based on ImageNet). In contrast, augmented data preprocessing introduced random horizontal flipping, random rotations (±15 degrees), and random resized cropping (scale between 0.8–1.0), thereby simulating observational variations and improving robustness.

All CNN models were trained using *PyTorch*. Training employed the Adam optimizer with an initial learning rate of 0.001, incorporating a learning rate scheduler to reduce the learning rate by half every 10 epochs. Models were trained for a maximum of 55 epochs, with checkpoints saved based on peak validation accuracy.

Redshift and Mass Prediction Using Bayesian Inference

Data Selection: The data used in this part of the investigation comes from the Buzzard V-1.0 Simulation, which generated realistic magnitude, uncertainty, mass, and redshift data for 111,172 based on parameters and methods used by the Rubin Observatory in Chile. The magnitude and magnitude uncertainty data come from astronomical measurements taken of light emitted by each galaxy using six different filter bands: *u* for ultraviolet, *g* for green, *r* for red, *i* for infrared, and *z* and *y*, representing different infrared bands. Also included is the redshift of each galaxy, representing by what factor light from the galaxy is stretched by the expansion of the universe as it travels from the galaxy to Earth. A larger redshift indicates that a galaxy is further away from Earth, and thus moving away from Earth at a higher velocity. Finally, the logarithm of the mass of the galaxy (in solar units) is included in the dataset. Because the logarithm of mass was included, both the raw mass data, and the exponentiated mass values, were experimented on in this analysis; the raw mass data turned out to have a much better fit, so it will be presented here.

Data Preparation: To capture both linear and non-linear relationships between the input magnitude data and the output redshift, it was decided to model both the mass and redshift using degree-2 interactions, including all possible cross terms of the *u*, *g*, *r*, *i*, *z*, and *y* parameters, resulting in the predicted redshift $\hat{r}$ being given by $\hat{r} = \sum_{<a,\,b>} W_{ab} m_a m_b$, where a and b represent each pairwise combination of filters and the constant term β (including combinations of the same filter), W_{ab} represents the weight of the given magnitude term, and m_a representing the magnitude of filter (or constant) a. Initializing a bias term allows for the establishment of a "baseline" galaxy redshift, from which information from the color magnitudes can be used to update the prior distribution iteratively. As such, the data was used to calculate every possible cross-term up to degree 2 (e.g. gz, $i\beta$, r^2) for each galaxy.[1] These cross-terms then served as the input for the linear Bayesian predictor.

[1]Although it looks like a cross term, when doing the actual data analysis, terms that included the constant β (e.g. $i\beta$) were simply just linear terms of the measurement of the other filter data (in this case, *i*). The term β^2 in the above sum is the true "bias" term, in traditional machine learning. It was just easier and cleaner to express the sum of quadratic, linear, bias, and cross-terms as a sum of pairwise cross-terms.

Bayesian Modeling: Because this investigation aimed to determine whether simpler statistical models can accomplish similar predictive power as more complex models, a multivariate Gaussian prior was assumed for each model weight, rather than the prior outlined in the Benítez paper. The polynomial-prepared input data was then split into an 80/20 training/test set to serve as the source of model training and testing. By using the polynomial-calculated magnitude data, the posterior distribution is calculated, thus outputting mean estimates for galaxy mass and redshift and also standard deviation estimates, which can be taken in tandem to produce a multivariate Gaussian posterior distribution for all input features. Although the uncertainty in each magnitude measurement and the posterior uncertainty aren't used explicitly to reach conclusions about the data, this process can easily be modified to reflect the magnitude uncertainty in the prior distribution and use the posterior distribution uncertainty to extend the analysis.

Machine Learning Results and Interpretation

Galaxy Morphology Classification Results

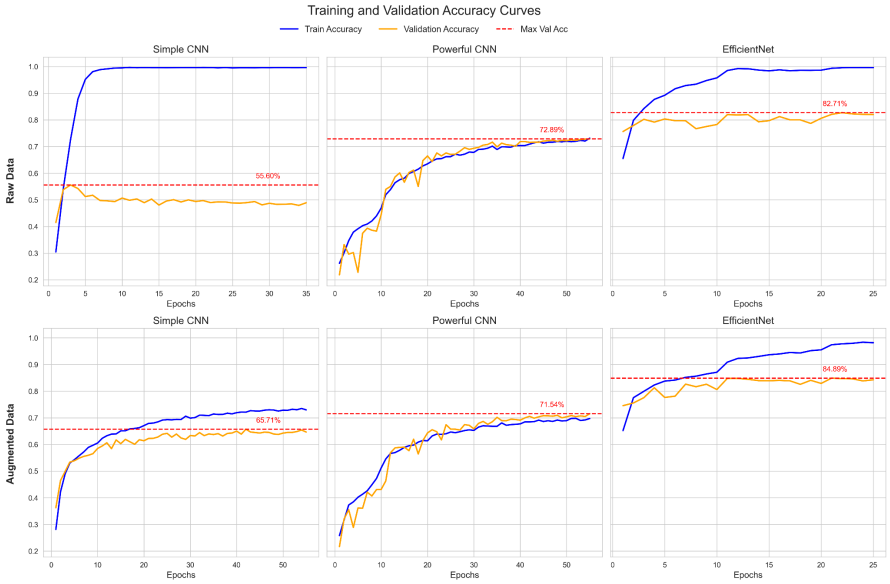

Figure 2: Training (blue) and validation (orange) accuracy curves for SimpleCNN, PowerfulCNN, and EfficientNet on both raw (top) and augmented (bottom) data. Dashed lines mark the highest validation accuracy achieved for each model.

Training and Validation Performance: Figure 2 presents the training and validation accuracy curves for the three convolutional neural network models (SimpleCNN, PowerfulCNN, and EfficientNet) under both raw (top row) and augmented (bottom row) data conditions. Several key trends emerge from this comparison that help explain the varying performance levels across models and training regimes.

One of the most evident effects is the beneficial impact of data augmentation on generalization. For all three models, the validation curves with augmented data (bottom row) exhibit reduced divergence from the training curves. This suggests that the application of image transformations improves the model's ability to generalize unseen data by effectively expanding the diversity of the training distribution. In contrast, the models trained on raw data often show wider gaps between training and validation performance, particularly for SimpleCNN, indicating potential overfitting.

Among the three models, PowerfulCNN stands out for its training stability. Its validation accuracy tracks remarkably closely with training accuracy in both raw and augmented settings, suggesting strong resistance to overfitting despite the model's greater depth and parameter count. This stability can be attributed to architectural choices: the use of batch normalization after every convolutional layer and the presence of dropout in the classifier likely contribute to more consistent generalization across epochs.

Improved Performance with Model Complexity: As expected, both model depth and representational power correlate with improved accuracy. The SimpleCNN serves as a strong but basic baseline; its performance is limited by architectural simplicity, as reflected in the lower validation accuracy (~55.6% on raw data, ~65.7% on augmented). In contrast, PowerfulCNN reaches significantly higher peak accuracies (~72.9% with raw data and 71.5% with augmentation), despite being trained from scratch. EfficientNet surpasses both, achieving the highest validation accuracies: 82.7% with raw data and 84.9% with augmentation.

EfficientNet's performance highlights the strength of transfer learning. Unlike the other models, it converges quickly, reaching strong accuracy within ~10 epochs, and maintains a high level of validation performance throughout training. This rapid and robust convergence underscores the advantage of initializing the model with pretrained weights from large-scale datasets such as ImageNet. By leveraging pretrained features, EfficientNet adapts quickly to the morphology classification task.

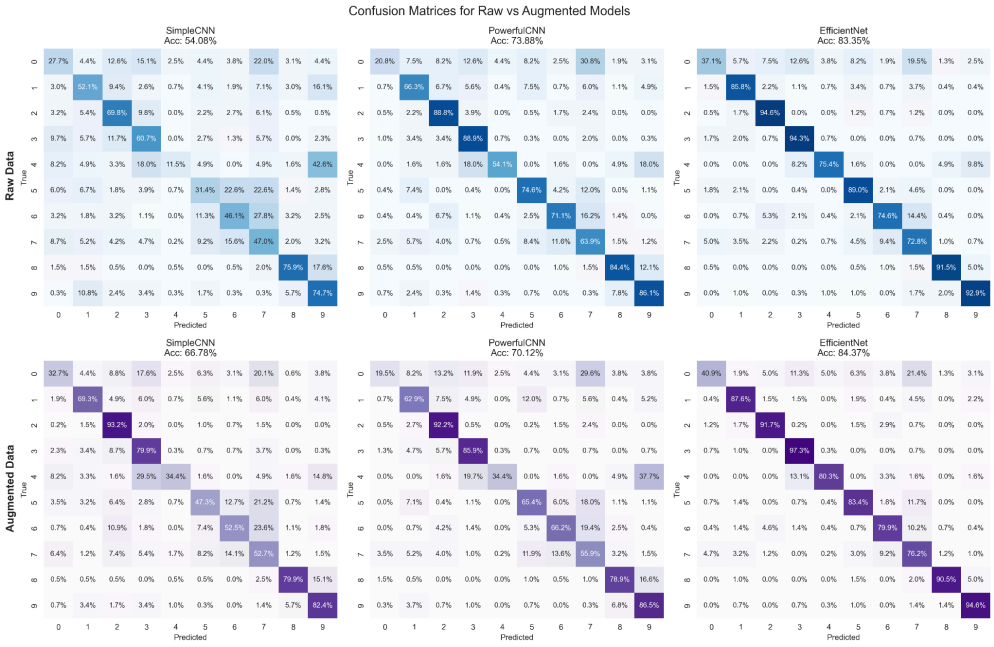

Figure 3: Confusion matrices illustrating class-by-class performance for SimpleCNN, PowerfulCNN, and EfficientNet under raw (top) and augmented (bottom) training. Higher (darker) diagonal values indicate more accurate predictions.

In the confusion matrices shown in Figure 3, each row corresponds to a true galaxy morphology class, while each column indicates the model's predicted class. Across both raw and augmented settings, we observe that data augmentation generally reduces off-diagonal errors, suggesting improved class-level generalization. Models trained on augmented data show notably fewer misclassifications in categories like merging or round smooth galaxies compared to their raw-data counterparts.

A particularly striking result is EfficientNet's ability to classify cigar-shaped smooth galaxies (Class 4), the least represented category in the dataset, more accurately than SimpleCNN or PowerfulCNN. Although there are only 334 images in Class 4, EfficientNet's transfer learning approach and overall advanced architecture appear to enable it to learn subtle morphological cues. In contrast, the simpler models show higher rates of confusion for this class, partly reflecting the difficulty of distinguishing cigar-shaped galaxies from other elongated or edge-on spiral forms.

Bayesian Inference Prediction Results

Output Architecture: After the Bayesian inference model was trained, each cross-term weight W_{ab} and covariance Σ_{ab} were determined, and because the model assumed a Gaussian prior, the galaxy training data updated each parameter weight to produce a Gaussian posterior as well. Given these weights and the multivariate linear regression sum identified in section II, the redshift and mass of each galaxy in the test set were estimated, with the prediction results being shown below against the true values of mass and redshift.

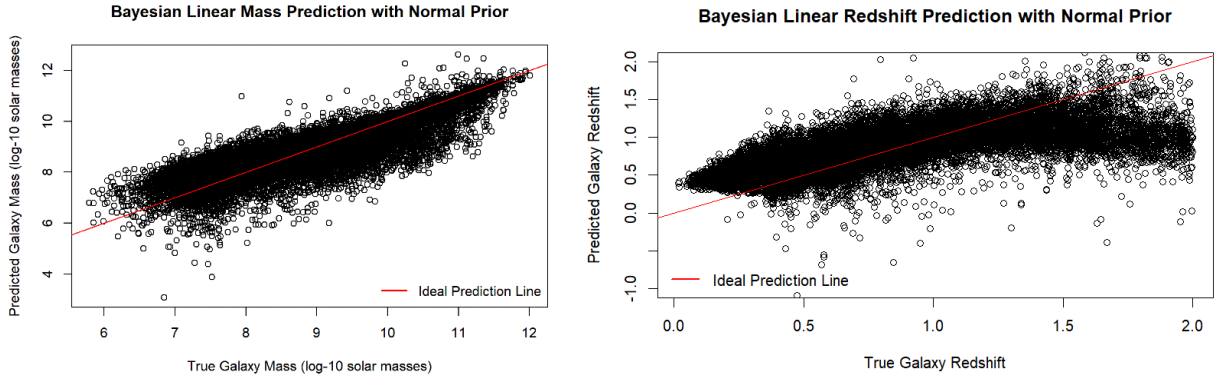

Figure 4: Scatter plots of the true and predicted mass and redshift of each of the more than 20,000 galaxies in the test set, along with the line y=x showing the positions of perfect predictions.

Posterior Weight Means: As specified above, the linear Bayesian model outputted a Gaussian posterior distribution for each of the input weights and their cross-terms. Because the original Benítez paper described a Bayesian approach for predicting the output distribution of redshift, most of the analysis was performed on the redshift data, with the mass prediction being included as an investigation for how generalizable this methodology is for Bayesian predictors of galaxy data. Thus, the following analysis focuses on the redshift predictor. The mean of the true bias, the weight of the β^2 term, ended up being 0.83, serving as a "baseline" redshift prediction for an arbitrary galaxy, from which the magnitude data from a given galaxy can be used to update this prediction. The parameter weight means varied in magnitude from 0.83 to 769.93. The large disparity in magnitudes between the parameter weight means and the β^2 weight indicates that perhaps there might have been some overfitting of the model to the training data, since the output prediction seems to vary wildly with small changes to the input values. The full matrix of mean posterior weights for the redshift analysis is shown in the plot below. The weight means of the posterior mass distribution showed even more variable results, with a variation in magnitude from 7.22 to 1508. Again, there may have been some issues with overfitting and lack of convergence in this model, and since these problems were common to both the mass and redshift predictors, there may have been some flaws or room for improvement in the model architecture.

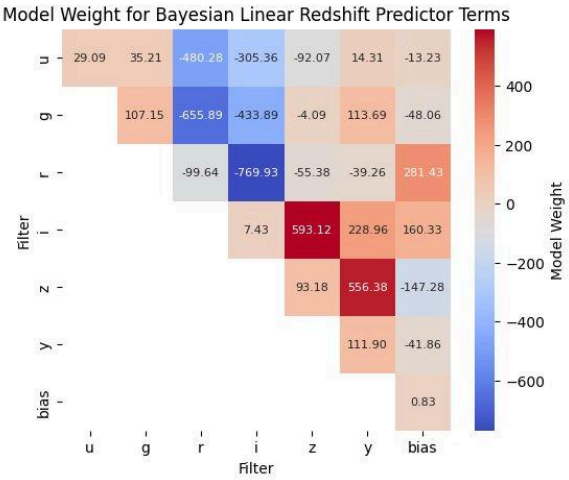

Figure 5: A heatmap matrix showing the mean values of the weights of each filter cross-term in the posterior redshift distribution, with the lower left triangle being omitted, as it is a duplicate of the upper right triangle.

Posterior Weight Covariance: The terms that seemed to matter the least were most terms involving the constant β, while all across the board, each cross-term that involved measurements from the infrared band i seemed to have the largest magnitude in mean weights. Most of the terms in the covariance matrix were relatively small (of magnitude less than 0.1), with some notable exceptions, especially $\Sigma_{ur,\,ur} = 895.270$ These high covariances, as well as the large posterior mean weights, suggest that perhaps there should be further investigation into the model's convergence to see how different hyperparameters could be selected or model architecture perfected to minimize the elements in the covariance matrix, reflecting a higher certainty in model weights. A similar story is seen in the mass covariance matrix, where most covariances have a magnitude < 0.1, while outliers such as $\Sigma_{gr,\,gr}$ and $\Sigma_{ui,\,ui}$ have magnitudes of almost 1000. It is important to note that most outlying covariance terms, both in the mass and redshift posteriors, are those that lie along the diagonal, implying that in the posterior distribution, most features have little covariance with one another, implying near independence between different cross-terms.

Model Performance: Despite the wild variation in mean posterior weight magnitudes and large variances in the dataset, looking at the results of Figure 4, in conjunction with statistics between the prediction and test set galaxies, tell a more promising story. Looking at the mass predictor, the correlation between the predicted mass set and test mass set is 0.872, and the mean squared error of the predictions is 0.302, or a normalized root mean squared error of 9.16%. This implies that the predicted and true galaxy masses have a moderate to strong linear correlation, but this relationship may not be that of equality (i.e. lying on the line y = x). On the other hand, the redshift predictor yielded a correlation between the prediction and test sets of 0.698 and a mean squared error of 0.08, or a normalized root mean squared error of 14.14%. Thus, the redshift predictor resulted in a slightly less accurate and slightly less linear relationship between the true and predicted data.

Conclusions and Future Work

Improving Morphology Classification: More Robust Architecture

This investigation consisted of a two-part machine-learning approach to the analysis of galaxy morphology and parameter estimation. Both parts were designed from a common inspiration: to infer properties of galaxies from their magnitude data in different transmission bands, using machine learning to speed and scale up what used to involve

manual observations and tedious calculations. The first strategy was an image-based classifier using different convolutional neural networks with increasing complexity, from a simple CNN model to a more complex transfer learning model. The results from this section of the project showed that model sophistication and transfer learning significantly improved classification accuracy, with EfficientNet achieving the highest performance out of the three network architectures compared. Data augmentation, including randomly cropping, rotating, and scaling further improved generalization across all CNN architectures, confirming the value of having a larger and more diverse training set for CNN models.

While the accuracy of each neural network was very high, especially for a 10-way classification system, there is some potential room for improvement. Most of the additions that were added when transitioning from SimpleCNN to PowerfulCNN, and then to EfficientNet were appropriate, as they are standard machine learning techniques used in a wide array of applications. But given more time and computation power, some additional features could be to explore more complex network architectures, such as ResNet, incorporating transfer learning from a more robust CNN like AlexNet, or even just making the network wider or deeper, up to a certain point. More accurate predictions can be obtained by adding even more data augmentation operations, specifically hue shifting; because galaxy color (and more specifically the magnitude measurements in different bands) seemed to play such an important role in determining galaxy classification, training a network on hue-shifted inputs (where the relative hue differences are still the same), could lead to a more well-trained network due to the increased size of the training data.

Improving Bayesian Predictive Power: More Informed Priors

The second part of this investigation involved a Bayesian linear regression model to predict galaxy mass and redshift using photometric magnitude data. Using second-degree polynomial interactions and Gaussian priors for the magnitude weights allowed expression of linear and non-linear relationships, and also allowed for a more intricate description of the distribution of model weights. While the model produced moderate-to-strong correlations and reasonable normalized root mean squared errors, large variances in the posterior weights and some high covariance values suggest potential issues with overfitting and lack of convergence. These observations suggest that certain assumptions of the Bayesian model, such as assuming a multivariate Gaussian prior and only degree-2 interactions, may be too simplistic, or just inappropriate, for capturing the full complexity of the data.

Along these lines, the analysis was generally appropriate given the scope of the investigation. In the spirit of the Jasche and Wandelt paper, these simplifications had some predictive power, but also had clear drawbacks. For one, the model struggled with high variances and sensitivity to small changes in the magnitude values, especially in higher-order terms, and may not generalize well to more test data without the implementation of regularization or more sophisticated prior distributions. Future work could explore integrating measurement uncertainties (which were given in the dataset) directly into the prior, a more detailed tuning of the prior hyperparameters, and actually implementing the Benítez prior, all of which could improve performance and interpretability.

Despite clear outlines for improvements for future investigations into the predictive properties of galaxy light spectra, the work done in this investigation show that there is power contained in the measurement of galactic magnitudes in different astronomical bands, and that machine learning certainly has its place in astronomy.

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.