Evaluation of Galaxy Morphology Classification with Machine Learning and Deep Learning

Yuhan You

School of Data Science, University of Virginia, Charlottesville, Virginia, United States of America
mxr9et@virginia.edu

Abstract. Through the development of technologies, astronomical imaging surveys have increased significantly, so that more galaxy images are taken than ever, which makes the traditional manual galaxy classification infeasible. For this reason, adopting automated machine learning techniques is important in replacing the traditional way of galaxy classification that requires cooperation with large-scale sky surveys. This research examines the performance of traditional and deep learning models through classifying galaxies from ten morphological features. The study used the Galaxy10 DECaLS dataset, which contained 17,736 colored galaxy images with labels. The dataset was first normalized and separated into training, validation, and testing subsets in proportions of 70, 15, and 15 percent. Then, the subsets of data were fed to the models. Traditional models like Support Vector Machine (SVM) with an RBF kernel and Random Forest with 500 estimators were applied to PCA-compressed features, and neural architectures like Multilayer Perceptron (MLP) and Convolutional Neural Network (CNN), were trained directly on images. The result indicates that the deep learning models substantially outperform the classical methods. CNN achieved the highest accuracies and best generalization to unseen data, which confirms the advantage of convolutional feature extraction over hand-engineered representations. However, deep learning models require large computing power and a dataset to reach an ideal high accuracy. Those findings show that the deep learning models show scalable, accurate morphological classification, which offers significant potential for future large-scale astronomical surveys.

Keywords: Galaxy Morphology Classification, Deep Learning in Astronomy, Convolutional Neural Networks (CNNs), Galaxy10 DECaLS Dataset, Machine Learning for Sky Surveys.

1 Introduction

In the early 20th century, people believed that the Milky Way was the entire observable universe, until astronomer Edwin Hubble discovered Cepheid variable stars to measure the distance of the true universe that expanding human's recognition of the scale of the cosmos [1]. The dark area besides the visible star hides billions of galaxies. The morphological features, such as spiral arms, bars, and elliptical shapes,

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_63

of those galaxies provide clues about a galaxy's age, composition, and interaction history [2,3]. With that information, scientists could have a better understanding of the formation of the universe. The classification of galaxies becomes crucial for the interpretation of galaxies.

For decades, astronomers classified these galaxies manually, which demanded enormous amounts of time and effort, because at the time, humans were better at distinguishing the structure of galaxies than computers. In the Galaxy Zoo project, thousands of volunteers gathered on the internet to classify tens of thousands of galaxy images [4]. Although this approach was successful, with the rise of large sky surveys like SDSS [5] and DESI [6], the amount of galaxy images has grown exponentially, making manual classification infeasible. This reality calls for awareness of the rise of data-driven astronomy and the importance of developing complicated machines and deep learning algorithms. Thus, scientists show interest in complicated machine learning and deep learning that automate morphological classification, which could reduce significant labor and time usage [7]. On the one hand, traditional algorithms such as Support Vector Machines (SVM) and Random Forests can capture global statistical features, while on the other, convolutional neural networks (CNNs) are superior at learning spatial hierarchies directly from images [7, 8].

This research explores the performance of different models when identifying ten classes of galaxies from the Galaxy10 DECals dataset [9]. Comparing the accuracy and efficiency of SVM, Random Forest, Multilayer Perceptron (MLP), and CNN models under identical conditions. The research aims to understand not only which model performs best but also how deep learning reshapes the ability to study the cosmos through data in this data-driven age.

2 Dataset and Preprocessing

2.1 Dataset

The experiments in this research adopt the Galaxy10 DECaLS dataset is provided and maintained by the AstroNN research group at the Max Planck Institute for Astronomy in collaboration with the Sloan Digital Sky Survey (SDSS) and the Dark Energy Spectroscopic Instrument (DESI) Legacy Imaging Surveys. AstroNN is an open-source astronomical machine learning initiative that curated and standardized the dataset for educational and research use. The Galaxy10 DECaLS is an enhanced version of the original Galaxy S10 dataset. The Galaxy10 DECaLS is composed of 17736 images with scale of 256x256 pixels colored galaxy images (g, r and z bands) categorized into 10 classes by shape. "The ten classes range from disturbed and merging galaxies to smooth, round shapes and edge-on disks. Each class is defined based on volunteer votes from the Galaxy Zoo project and uses images drawn from the DESI Legacy Imaging Surveys (DECaLS) to provide improved resolution and image quality" (astronn.readthedocs.io). In AstroNN's website, the classification label of the morphologies are describe as the following: Disturbed (1081 images), Merging (1853 images), Round Smooth (2645 images), In-between Round Smooth (2027

images), Cigar Shaped Smooth (334 images), Barred Spiral (2043 images), Unbarred Tight Spiral (1829 images), Unbarred Loose Spiral (2628 images), Edge-on without Bulge (2628 images), and Edge-on with Bulge (1873 images). Noticed that Cigar Shaped Smooth Galaxies have significantly less distribution, which might cause potential bias. The Galaxy10 DECaLS dataset is well organized and friendly for machine learning purposes.

2.2 Data Preprocessing

For data preprocessing due to the limitation of computing power, the images are resized and normalized to ensure stability and consistency across all the models. All galaxy images are downscaled into 64x64x3 dimensions to reduce the load of the computer while making sure sufficient morphological features for the model to interpret. Meanwhile, pixel intensity values are converted to floating-point, which are scaled into the range [0,1] and divided by 255.0. The class labels are checked and standardized: the minimum label value is shifted to 0 if necessary, and the unique label sets for training and test splits are compared to ensure they match exactly (10 classes).

Then, the dataset is separated into training, validation, and testing subsets. This research adopts a strategy that splits 70% for training, 15% for validation, and 15% for testing. Additionally, for the machine learning models (SVM, Random Forest), dimensionality reduction is applied: the flattened image vectors are processed via Principal Component Analysis (PCA) that preserves approximately 85% of variance, in order to reduce feature dimension and improve classifier performance.

All models receive the same input of training, validation, and testing subsets of the Galaxy10 DECaLS dataset. The preprocessing pipeline ensures that all models receive the same input schema, which secures the fairness of the comparison between classical and deep-learning models' approaches.

3 Mythology

3.1 Machine Learning Models

Support Vector Machine (SVM): The classifier is built with a radial basis function (RBF) kernel, which allows nonlinear decision boundaries in feature space. Before fitting the SVM, PCA is applied to the standardized pixel vectors, retaining approximately 85% of the total variance.

Random Forest (RF): The classifier was implemented using scikit-learn, and configured with 500 trees, unlimited depth, and full parallelization across CPU cores. Before fitting the SVM, PCA is applied to the standardized pixel vectors, retaining approximately 85% of the total variance.

3.2 Deep Learning Models

Multilayer Perceptron (MLP): The MLP baseline model consisted of an input flattening layer followed by two fully connected layers (512 and 256 neurons) with ReLU activations and 0.3 dropout. The final dense layer used a SoftMax activation

across 10 classes and was trained with 20 epochs using the Adam optimizer, sparse categorical cross-entropy, and a batch size of 64.

Convolutional Neural Network (CNN): The CNN model was built with three convolutional blocks (32, 64, 128 filters) and each followed by Batch Normalization and Max Pooling layers. The extracted features were flattened and passed through a dense layer (256 units) with 0.4 dropout before the final SoftMax layer. The network was trained for 25 epochs using the Adam optimizer and early stopping to prevent overfitting.

All models were trained in VS Code using TensorFlow 2.16 on a CPU (Intel i9-14900K). The experiments were executed inside an isolated virtual environment to ensure consistent dependencies across runs.

4 Experimental Results

4.1 Evaluation Metrics

This research used the overall accuracy of the classification as the primary metric for the evaluation of model performance. Additionally, the complementation of precision, recall, and F1-score are considered to assess per-class behavior. Confusion matrices were created for the evaluation of each model as well in order to visualize inter-class classification accuracy and misclassification patterns.

4.2 Quantitative Results

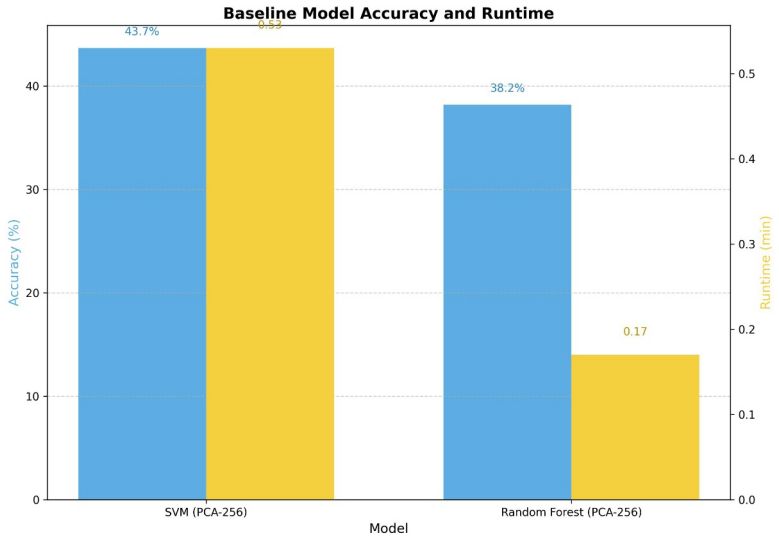

Fig. 1. Accuracy and Runtime (Picture credit: Original)

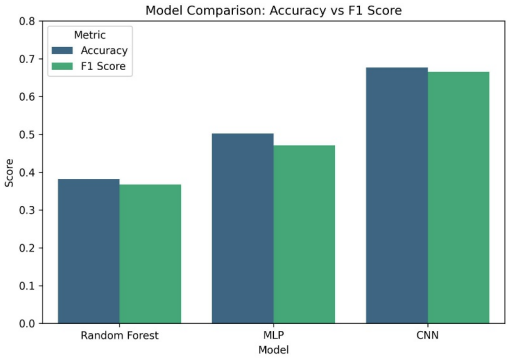

Fig. 2. Accuracy vs F1 Score (Picture credit: Original)

Figure 1 shows the accuracy and runtime of both traditional models. The SVM (PCA-256) achieved 43.7% accuracy and a runtime of 0.53 minutes, and the Random Forest (PCA-256) yielded 38.2% accuracy with a shorter runtime of 0.17 minutes. This graph demonstrates the traditional models lacking the ability to interpret pictures due to their simplicity, which can be proved by their runtime.

Figure 2 summarizes the performance of 2 deep learning models and compares them to a traditional model (Random Forest). CNN outperforms all other models and achieves approximately 68% accuracy and 0.67 F1-score, which shows its superior ability to generalize and balance precision and recall across all galaxy classes. MLP on the other hand The MLP achieves ~50% accuracy and 0.47 F1-score, reflecting moderate performance due to its lack of spatial feature extraction but strong generalization stability. Last, The Random Forest performs the weakest with ~38% accuracy and 0.36 F1-score, consistent with its limited capability to model complex nonlinear visual features after PCA compression.

4.3 Train and Validation Curves

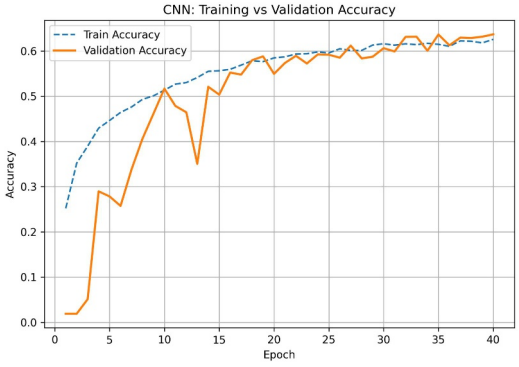

Fig. 3. Training vs Validation Accuracy (Picture credit: Original)

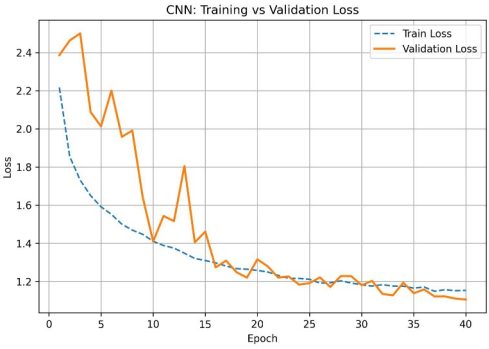

Fig. 4. Training vs Validation Loss (Picture credit: Original)

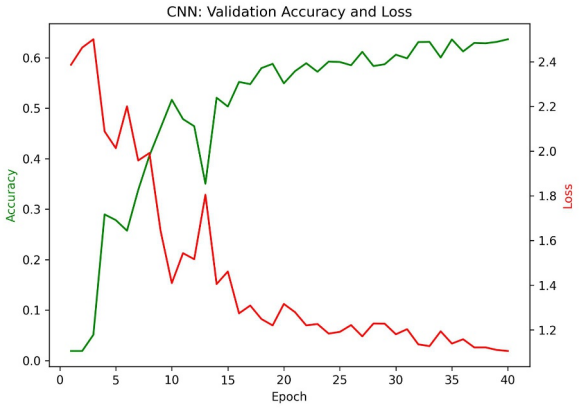

Fig. 5. Dual-Axis Validation Accuracy and Loss Plot (Picture credit: Original)

Figure 3 shows the rapid improvement of the CNN model in the early epoch. As the validation accuracy rose in the first 10 epochs, the accuracy stabilized around 65% by the 40th epoch, and, moreover, the gap between training and validation accuracy narrowed as training progressed. This highlights the successful regularization and proper learning rate scheduling of the CNN model. In Figure 4, both training and validation losses decrease substantially. While the validation loss shows some oscillation in early epochs due to data augmentation randomness, the curves converge below 1.2 loss and consistently achieve a test accuracy of around 68%. Last, the dual-axis plot (Figure 5) visually reinforces the inverse correlation between validation accuracy (green) and loss (red).

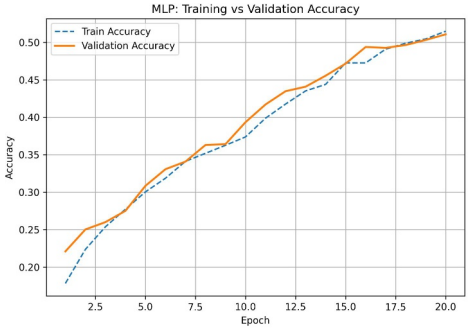

Fig. 6. MLP: Training vs. Validation Loss (Picture credit: Original)

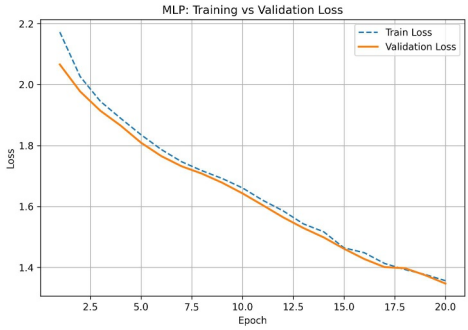

Fig. 7. MLP: Training vs Validation Loss (Picture credit: Original)

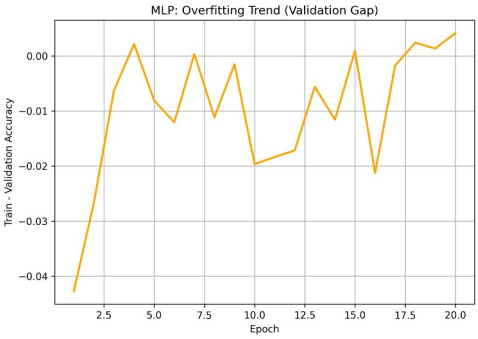

Fig. 8. MLP: Overfitting Trend (Validation Gap) (Picture credit: Original)

Figure 6 shows a steady increase in both training and validation accuracy of the MPL model around 20 epochs and later converges at around 50% accuracy. The two curves remain consistent and close throughout the training, indicating generalization. In Figure 7, the training and validation loss follows the inverse shape of training and

validation accuracy, which declines smoothly from roughly 2.1 to 1.4. The lack of divergence between them implies that the model maintains stable optimization and avoids memorization of the training data. Last, in Figure 8, the validation gap fluctuates near zero, which shows no consistent increase over epochs. This means the MLP does not overfit.

4.4 Confusion Matrix Analysis

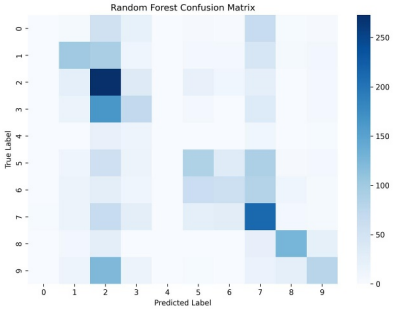

Fig. 9. Random Forest Confusion Matrix (Picture credit: Original)

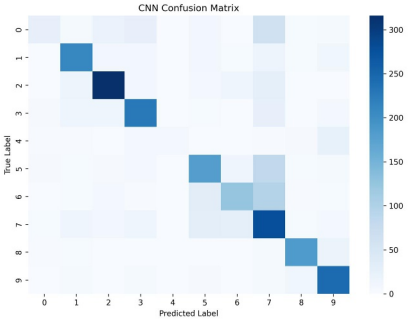

Fig. 10. CNN Confusion Matrix (Picture credit: Original)

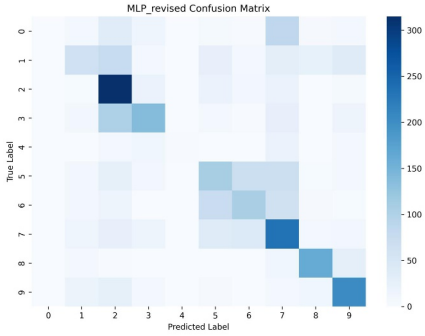

Fig. 11. MLP_revised Confusion Matrix (Picture credit: Original)

The Random Forest confusion matrix in figure 9 reveals inconsistent class separation, with fewer strong diagonal regions and several classes overlapping substantially. This pattern suggests that while RF captures general trends in the reduced PCA feature space, it is limited by compressed dimensional representation and lacks spatial locality awareness. The overall accuracy aligns with this weaker class distinction.

In figure 10, Convolutional Neural Network the CNN confusion matrix demonstrates strong diagonal dominance. This pattern indicates that CNN effectively learned spatial and texture-based representations. Minor misclassifications occur only between visually similar categories, which reflects the model's ability to capture the hierarchical image features of galaxies. Overall, CNN shows the highest inter-class separation among all models. Multilayer Perceptron MLP confusion matrix shows less concentrated diagonals and broader off-diagonal distributions in figure 11. This indicates that while the MLP learns some global patterns, it struggles with fine-grained spatial structures due to the lack of convolutional feature extraction. The confusion is most noticeable between classes with subtle differences in shape or brightness. The MLP still captures partial structure, but it lacks the representational depth of the CNN.

5 Discussion

5.1 Summary

This study investigated the performance of both traditional and deep learning approaches for automated galaxy morphology classification using the Galaxy10 DECaLS dataset through the classification accuracy of the models. Traditional models, including Support Vector Machine (SVM) and Random Forest (RF), achieved low accuracy after PCA-based dimensionality reduction because they struggled to capture the spatial complexity inherent in galaxy structures. On the other hand, the Convolutional Neural Network (CNN) and fine-tuned EfficientNetB0 demonstrated significantly higher accuracy and better generalization because of better recognition of the galaxy structure. These findings are consistent with prior research by Domínguez Sánchez et al. [10], who used deep convolutional networks to classify over 670,000 SDSS galaxies with expert-level accuracy, and Barchi et al. [7], who conducted a direct comparison of traditional and deep learning models and found that deep architectures achieved up to 99% accuracy in two-class galaxy separation tasks. The results of this study further confirm that convolutional and transfer-learning models outperform traditional methods.

5.2 Interpretation

The research created, tested, and evaluated both the traditional and deep learning models, including SVM, Random Forest, MLP, CNN, and EfficientNetB0. The results showed that the deep learning models outweigh the performance of the

traditional models, with the best performed traditional model having an accuracy of 43.67% and the deep learning model has an accuracy of 63.67%. However, the deep learning models required much more time to train due to their complexity, 100 to 200 times longer, yet with a good computing resource, this problem could be addressed. In the case of image classification, the deep learning models' accuracy matters more than their efficiency, which makes them more suitable for such a task.

5.3 Limitations

The accuracy is still far off from the ideal accuracy of 97.5% or higher, due to the following reasons. First, the lack of computing power really limits the depth of this research, as each image is reduced to 64x64, losing much of its features from 256x256, which could enhance accuracy. Second, although the dataset of 17000 images seems a lot, the scale of the dataset is not enough to secure higher accuracy, for example, in Barchi's experiment 60000 images used in the study. However, the current equipment is not good for providing a stable environment to run such an amount of data. Lastly, the model adoptions are on the easier side of deep learning models, the advanced models with multiple layers are rejected under the same reason as the previous two limitations.

5.4 Future Work

The ideal future work is to conduct experiments with access to high-end GPUs, which means a boost in computing power. Thus, larger datasets and complicated models could be adopted for advanced study. Meanwhile, unsupervised models are of particular interest for future related works. Ultimately, combining larger datasets, advanced architectures, and unsupervised techniques under a more powerful computing infrastructure could greatly enhance both accuracy and interpretability, pushing automated galaxy morphology classification toward scientific-grade reliability.

6 Conclusion

In short summary, the research compared the performance of traditional and deep learning models through their accuracy and F1-score of classifying the ten galaxy morphological features provided by the Galaxy10 DECaLS dataset. The results of this study show that traditional models, SVM, and Random Forest lack the ability to capture 2D spatial relationships underlying the images, which leads to poor performance. On the other hand, CNN and MPL's superior performance is attributed to their recognition of 2D features of each morphology in the images. Despite this success, the study remains constrained by the dataset size and class imbalance, and the computing power that restricts the model depth. Thus, the future work should focus on increasing the scale of the experiment with larger datasets and advanced GPU acceleration and AI techniques, so that the performance of each model can be enhanced. Overall, the findings highlight the transformative role of deep learning in advancing automated galaxy classification and observational cosmology in the future.

References

1. Hubble, E.: The Realm of the Nebulae. Yale University Press, New Haven, CT (1936)
2. Conselice, C.J.: The evolution of galaxy structure over cosmic time. Ann. Rev. Astron. Astrophys. 52, 291–337 (2014)
3. Blanton, M.R., Moustakas, J.: Physical properties and environments of nearby galaxies. Ann. Rev. Astron. Astrophys. 47, 159–210 (2009)
4. Walmsley, M., Géron, T., Kruk, S., Scaife, A.M.M., Lintott, C., Masters, K.L., Dawson, J.M., Dickinson, H., Fortson, L., Garland, I.L., et al.: Galaxy Zoo DESI: Detailed morphology measurements for 8.7 million galaxies in the DESI Legacy Imaging Surveys. Mon. Not. R. Astron. Soc. 526(3), 4768–4786 (2023)
5. Eisenstein, D.J., Weinberg, D.H., Agol, E., Aihara, H., Allende Prieto, C., Anderson, S.F., et al.: SDSS-III: Massive spectroscopic surveys of the distant universe, the Milky Way, and extra-solar planetary systems. Astron. J. 142(3), 72 (2011).
6. DESI Collaboration: Overview of the Dark Energy Spectroscopic Instrument (DESI). Astron. J. 165(6), 259 (2023)
7. Barchi, P.H., de Carvalho, R.R., Rosa, R.R., Sautter, R.A., Soares-Santos, M., Marques, B.A.D., Clua, E., Gonçalves, T.S., de Sá-Freitas, C., Moura, T.C.: Machine and deep learning applied to galaxy morphology – A comparative study. Astron. Comput. 30, 100334 (2020)
8. Song, S., Liu, B., Teng, F., Li, T.: Self-supervised contrastive learning for implicit collaborative filtering. Eng. Appl. Artif. Intell. 139, 109563 (2025)
9. Hausen, R., Robertson, B.E.: Morphen: Galaxy morphology prediction with deep learning. Astrophys. J. Suppl. Ser.248(1), 20 (2020)
10. Domínguez Sánchez, H., Huertas-Company, M., Bernardi, M., Tuccillo, D., Fischer, J.L.: Improving galaxy morphologies for SDSS with deep learning. Mon. Not. R. Astron. Soc. 476(3), 3661–3676 (2018)