Scalable Galaxy Morphology Classification through Hybrid Handcrafted and Deep Learning Approaches

Murilo L. C. Neves[1], **Maria Eduarda M. Policante**[1], **Valéria D. Feltrim**[1], **Yandre M. G. Costa**[1]

[1] Departamento de Informática
Universidade Estadual de Maringá
Maringá, PR, Brazil

Abstract. *The rapid growth of astronomical surveys has transformed galaxy morphology classification into a large-scale data analysis problem that requires automated and scalable machine learning solutions. In this work, we evaluate handcrafted and deep learning approaches using a balanced subset of the Galaxy Zoo 2 dataset containing seven morphological classes. We first investigate the impact of a multi-step preprocessing pipeline and extract handcrafted descriptors capturing textural, morphological, intensity, and structural properties of galaxies. These descriptors are evaluated using classical machine learning models, including SVM, MLP, Random Forest, and KNN. We then assess convolutional neural networks (AlexNet, ResNet50, MobileNetV2, and EfficientNetB0) using transfer learning on both preprocessed and raw centrally cropped images. Finally, we explore a hybrid ensemble combining handcrafted and deep representations. The best configuration, based on a weighted soft-voting scheme, achieved an F1-score of 0.929. These results indicate that handcrafted descriptors and deep representations capture complementary information, and that their integration can improve scalable scientific image analysis in data-intensive research environments.*

1. Introduction

In recent decades, astronomy has become an increasingly data-intensive science. Large-scale digital sky surveys continuously generate massive volumes of high-resolution images, turning galaxy morphology classification into a challenging large-scale data analysis task. Although traditional visual schemes such as the Hubble Sequence [Hubble 1926] remain conceptually important, manual classification is no longer feasible at the scale required by contemporary observational astronomy.

Citizen-science initiatives such as Galaxy Zoo [Fortson et al. 2011] and Galaxy Zoo 2 [Willett et al. 2013] have partially addressed this challenge by enabling volunteers to contribute to galaxy morphology labeling. These efforts have produced valuable annotated datasets and significantly advanced the field. However, they still depend on human effort and therefore face limitations in scalability as astronomical data acquisition continues to accelerate.

In this scenario, machine learning has emerged as a key enabler of automated and scalable scientific analysis. Both traditional machine learning methods and deep learning models, especially Convolutional Neural Networks (CNNs), have been investigated for galaxy morphology classification with promising results [Kalvankar et al. 2020, Zhu et al. 2019]. More broadly, such approaches reflect the ongoing digital transformation of scientific practice, in which intelligent systems play an essential role in converting large volumes of raw observational data into structured and actionable knowledge.

In this work, we investigate the effectiveness of classical and deep learning approaches for galaxy morphology classification using a balanced subset of the Galaxy Zoo 2 (GZ2) dataset. More specifically, we assess the impact of a multi-step preprocessing pipeline, evaluate handcrafted descriptors combined with classical machine learning algorithms, and compare them with transfer-learning approaches based on AlexNet, ResNet50, MobileNetV2, and EfficientNetB0. We also examine whether combining handcrafted and deep representations can improve classification performance in a seven-class morphological setting.

The remainder of this paper is organized as follows. Section 2 presents the related works. Section 3 describes the dataset. Section 4 presents the methodology, including preprocessing, feature extraction, and classification models. Section 5 reports and discusses the experimental results. Finally, Section 6 concludes the paper and outlines directions for future work.

2. Related Works

The necessity that visual frameworks like the Hubble Sequence [Hubble 1926] have of human observation leads to a challenge in a scenario where large amounts of data need to be dealt with. While citizen sciente initiatives like Galaxy Zoo [Fortson et al. 2011, Willett et al. 2013] help provide essential human-labeled data, this remains unscalable for upcoming surveys and data volume growth.

In an effort to automate this classification, some studies applied handcrafted features as descriptors for ML, such as [Shamir 2009], which uses automatic feature relevance determination to select the top 15% most relevant features out of 2873 computed, and [Ntwaetsile and Geach 2021], which uses Haralick features. Moreover, CNN shifted the paradigm by automatically learning representations directly from pixels, leading to studies such as [Zhu et al. 2019], which used CNNs to perform a classification task with five classes, and [Kalvankar et al. 2020], which expanded on the former study by increasing the number of classes to seven and exploring EfficientNet architectures.

Nevertheless, pure deep learning models can overlook specific geometric constraints that handcrafted descriptors explictly capture, at the same time that handcrafted descriptors may not capture nuanced details in the same way as CNNs. This leaves an opportunity to explore hybrid strategies and its effectiveness, which our work aims to do by evaluating whether an ensemble of handcrafted shallow models and CNNs can enhance robustness in a multi-class scenario.

3. Dataset

The dataset used in this study is derived from the GZ2 image release and its corresponding image-mapping dataset available on Kaggle [Trickz 2020]. The GZ2 dataset contains 243,437 RGB images with resolution 424×424. Each galaxy is associated with vote proportions obtained through the question-based annotation scheme described in [Willett et al. 2013]. The mapping dataset contains 355,991 entries linking object IDs (unique identifiers) to image IDs, since the GZ2 images are indexed according to their position in the dataset.

The vote proportions were used to assign each image to one of seven morphological classes following the criteria defined by [Kalvankar et al. 2020]. The considered classes were: completely round smooth, in-between smooth, cigar-shaped smooth, lenticular, barred spiral, unbarred spiral, and irregular, which were encoded with labels from 0 to 6, respectively. Examples of galaxies belonging to each class are shown in Figure 1.

The GZ2 images were first paired with the image-mapping dataset. Samples with no correspondence between image ID and object ID were discarded. Additionally, some images did not meet the criteria required to be assigned to any of the seven classes based on the vote proportions, and were also removed from the dataset. After this filtering process, 169,279 labeled samples remained, with the class distribution presented in Table 1. The complete filtering pipeline is illustrated in Figure 2.

Because the dataset exhibits a strong class imbalance, we adopted an undersampling strategy to obtain a balanced dataset and prevent class bias, ensuring that metrics reflect predictive capability rather than frequency distribution. Specifically, we randomly selected 5,199 images from each class, corresponding to the number of samples in the minority class. This resulted in a final dataset of 36,393 images used for training and evaluation, as summarized in Table 1.

Table 1. Class distribution before and after the undersampling procedure.

ID	Morphological Class	Original Samples	After Undersampling
0	Completely round smooth	42,778	5,199
1	In-between smooth	46,676	5,199
2	Cigar-shaped smooth	5,199	5,199
3	Lenticular	20,368	5,199
4	Barred spiral	5,297	5,199
5	Unbarred spiral	27,874	5,199
6	Irregular	21,087	5,199
Total		169,279	36,393

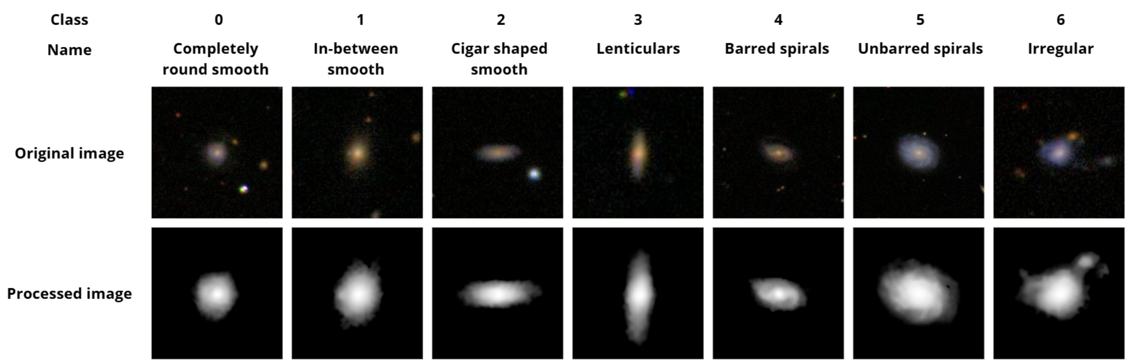

Figure 1. Examples of galaxies from the dataset

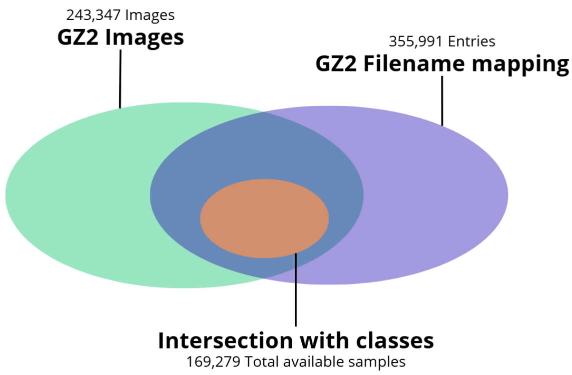

Figure 2. Intersection of labels, classes and images in the datasets.

4. Method

This section describes the methodological pipeline adopted in this study, including the preprocessing steps, handcrafted feature extraction, and classification approaches based on both classical machine learning and deep learning models.

4.1. Image preprocessing

The preprocessing pipeline applied to the raw images is illustrated in Figure 3 and consists of the following steps:

1. **Central cropping**: The input image is cropped to retain the central $50\% \times 50\%$ region (from $1/4$ to $3/4$ along each axis), focusing on the galaxy. This strategy follows the approach adopted in [Kalvankar et al. 2020] and [Zhu et al. 2019].
2. **Grayscale conversion**: The cropped RGB image is converted to grayscale.
3. **Denoising**: Non-local means denoising is applied to reduce noise while preserving structural details.
4. **Contrast enhancement (mask preparation)**: CLAHE (Contrast-Limited Adaptive Histogram Equalization) is applied with parameters clipLimit $= 3.0$ and tileGridSize $= (8, 8)$ to generate a high-contrast version used for mask extraction.
5. **Gaussian blur**: The enhanced image is smoothed using a 5×5 Gaussian kernel to facilitate segmentation.
6. **Otsu thresholding**: Otsu's method is used to convert the image into a binary mask.
7. **Morphological opening**: A 3×3 kernel is applied to perform morphological opening followed by dilation, removing small artifacts and closing minor gaps, which are sometimes present in telescopic imagery.
8. **Central connected component**: The connected component located at the center of the image is retained, while other components are discarded.
9. **Mask application**: The resulting mask is applied to the denoised image.
10. **Strong contrast enhancement**: A second CLAHE operation (clipLimit $= 5.0$, tileGridSize $= (4, 4)$) is applied to the masked region.
11. **Min–max scaling**: Pixel values inside the mask are rescaled to $[0, 255]$, while pixels outside the mask are set to zero.

After preprocessing, each image is represented as a 212×212 grayscale image centered on the galaxy. Although this procedure effectively isolates the main object in most cases, some images may still contain artifacts when multiple galaxies appear in close proximity.

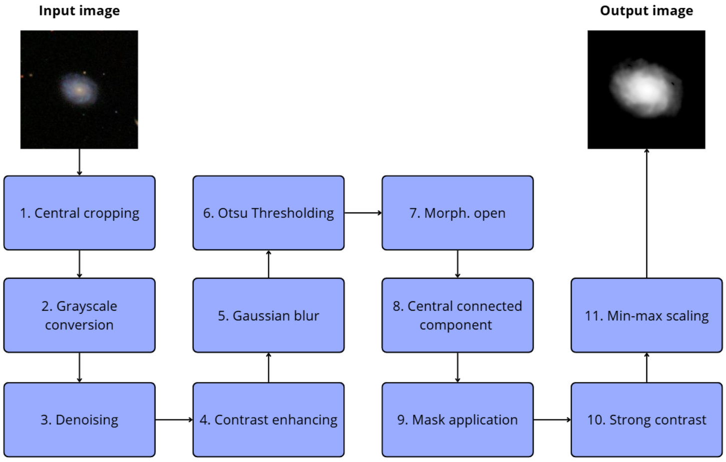

Figure 3. Preprocessing pipeline.

4.2. Handcrafted features

To provide inputs for classical machine learning algorithms, several handcrafted descriptors were extracted from the preprocessed images. These include histogram-based descriptors (HOG and LPQ) as well as a set of statistical and morphological descriptors combined into a Mixed Feature Vector (MFV).

Local Phase Quantization (LPQ) is a texture descriptor based on the blur-invariant properties of the Fourier phase spectrum [Ojansivu and Heikkilä 2008]. The method captures local phase information within small image windows and is known to be robust to blur and illumination variations. We computed LPQ using a window size of three, frequency estimation equal to 1.0, and decorrelation enabled.

Furthermore, Histogram of Oriented Gradients (HOG) is a widely used descriptor that captures local shape and appearance by computing histograms of gradient orientations [Dalal and Triggs 2005]. In this work, HOG features were extracted using eight orientations, a 32×32 cell size, and one cell per block.

4.2.1. MFV

The Mixed Feature Vector (MFV) aggregates several handcrafted descriptors into a single representation. The extracted features are grouped into four categories: textural features, morphological and geometric features, intensity and light profile features, and structural or frequency-domain features. In total, the MFV representation contains 21 numerical features.

1. **Textural features**: Gray-Level Co-occurrence Matrix (GLCM) statistics were computed using a one-pixel distance at $0°$. The extracted features include contrast, dissimilarity, homogeneity, energy (angular second moment), correlation, and Shannon entropy.
2. **Morphological and geometric features**: These descriptors were computed from the largest connected component located near the image center and include area, perimeter, eccentricity, orientation, solidity, axis ratio, circularity ($4\pi \cdot \text{Area}/\text{Perimeter}^2$), compactness ($\text{Area}/\text{Perimeter}^2$), and centroid offset relative to the image center.

3. **Intensity and light profile features**: These include the radial intensity standard deviation and the concentration index defined as $5\log_{10}(r_{80}/r_{20})$, where r_{20} and r_{80} correspond to the radii containing 20% and 80% of the total flux.
4. **Structural and frequency-domain features**: These descriptors include the asymmetry score (difference between the image and its $180°$ rotation), the box-counting fractal dimension, and the ratio of high-frequency energy in the Fourier spectrum.

4.3. Classification

The classification stage evaluates both handcrafted feature-based models and CNN arquitectures using transfer learning. All experiments used a $90/10$ train–test split following a holdout, which was chosen over cross-validation due to the large dataset size and the computational constraints of deep learning techniques.

Experiments were conducted using two types of input images: (i) preprocessed images obtained from the pipeline described earlier, and (ii) raw images where only central cropping was applied. This allows a direct comparison with previous studies such as [Kalvankar et al. 2020].

A general overview of the methodological pipeline is presented in Figure 4.

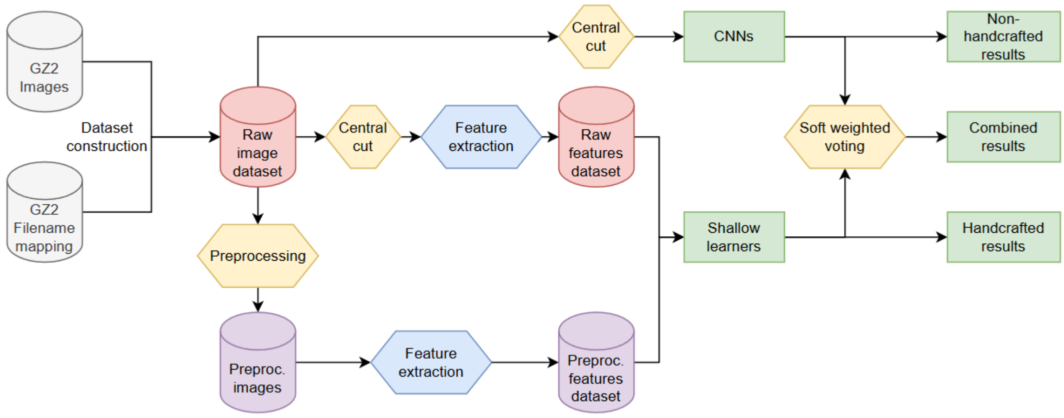

Figure 4. General view of our methodology.

4.3.1. Handcrafted features with shallow learners

Four classical machine learning algorithms were evaluated using the handcrafted descriptors: K-Nearest Neighbours (KNN), Random Forest (RF), Multi-Layer Perceptron (MLP), and Support Vector Machine (SVM). Hyperparameters for all models were selected through grid search.

Different feature configurations were explored, including individual descriptors, early fusion through feature concatenation, and late fusion through model combination. Features were extracted from both preprocessed images and raw centrally cropped images, as well as from the concatenation of both representations. The shallow learners were used with the Python library Scikit-Learn [Pedregosa et al. 2011].

4.4. Deep representations and architectures

For the deep learning approach, four widely used CNN architectures were evaluated: AlexNet [Krizhevsky et al. 2012], ResNet50 [He et al. 2016], MobileNetV2 [Sandler et al. 2018], and EfficientNetB0 [Tan and Le 2019]. All models were initialized with weights pre-trained on the ImageNet dataset [Deng et al. 2009], and only the final classification layers were retrained for the seven-class galaxy classification task. All tests were conducted on a laptop with a GTX 1650 video card and 4GB of VRAM.

To reduce overfitting and account for the arbitrary orientation of galaxies, data augmentation was applied during training. The augmentation strategy included random horizontal and vertical flips as well as random rotations within the interval $[-\frac{\pi}{2}, \frac{\pi}{2}]$.

5. Results

This section presents the results obtained with the handcrafted-based approach, followed by the CNN-based models and the hybrid strategy combining both representations.

5.1. Handcrafted features

The results obtained using handcrafted descriptors indicate that the Mixed Feature Vector (MFV) consistently provides the best performance among the evaluated feature sets. LPQ showed limited performance, while HOG yielded moderate improvements when combined with other descriptors.

When features were extracted from raw centrally cropped images, the best performance was obtained using the combination of HOG and MFV (F1 = 0.532). In contrast, when using the preprocessed images, the MFV alone achieved the best results (F1 = 0.633), as shown in Table 2.

In both cases, the best-performing classifier was the MLP. These results suggest that the preprocessing pipeline enhances morphological and structural descriptors, improving the effectiveness of classical machine learning models. However, the improved performance obtained by combining HOG and MFV in the raw-image setting indicates that these descriptors may capture complementary information that is partially reduced during preprocessing.

Despite the improvements obtained with the preprocessing pipeline, the concatenation of MFVs extracted from both raw and preprocessed images produced the best handcrafted results overall (Table 3). In this configuration, the MLP achieved F1 = 0.651, slightly outperforming SVM. This result suggests that combining representations from different preprocessing conditions preserves complementary discriminative information.

Finally, the best ensemble configuration using handcrafted features (Table 4) was obtained through soft voting between Random Forest and MLP using the concatenated MFVs, achieving F1 = 0.664. Although this represents an improvement over individual classifiers, the performance remains considerably lower than that obtained by the deep learning models.

5.2. Deep representations

The performance of the CNN architectures using preprocessed images is presented in Table 5. All networks achieved similar performance, with accuracy ranging from 78.3% to 81.5%. MobileNetV2 and ResNet50 produced the best results in this configuration.

Table 2. Results with different features from raw cut images and algorithms.

	Raw cut images				Preprocessed			
SVM	Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.
LPQ	0.212	0.200	0.214	0.212	0.287	0.266	0.265	0.287
HOG	0.349	0.315	0.354	0.349	0.501	0.488	0.488	0.501
MFV	0.513	0.513	0.515	0.513	0.631	0.625	0.629	0.631
HOG + MFV	0.526	0.523	0.532	0.526	0.601	0.592	0.600	0.601
All	0.517	0.516	0.521	0.517	0.580	0.571	0.577	0.580
MLP	Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.
LPQ	0.176	0.175	0.175	0.176	0.245	0.246	0.246	0.245
HOG	0.364	0.315	0.354	0.349	0.478	0.468	0.469	0.478
MFV	0.516	0.515	0.517	0.515	**0.633**	**0.626**	**0.629**	**0.633**
HOG + MFV	**0.534**	**0.532**	**0.533**	**0.534**	0.573	0.570	0.569	0.573
All	0.418	0.418	0.420	0.418	0.487	0.484	0.482	0.487
RF	Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.
LPQ	0.206	0.198	0.201	0.206	0.298	0.277	0.276	0.298
HOG	0.352	0.333	0.352	0.337	0.489	0.472	0.473	0.489
MFV	0.450	0.448	0.449	0.450	0.602	0.595	0.597	0.602
HOG + MFV	0.515	0.512	0.513	0.515	0.601	0.587	0.593	0.597
All	0.492	0.488	0.490	0.492	0.576	0.564	0.570	0.576
KNN	Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.
LPQ	0.165	0.171	0.165	0.154	0.233	0.201	0.230	0.233
HOG	0.325	0.301	0.315	0.322	0.426	0.388	0.424	0.426
MFV	0.404	0.404	0.415	0.404	0.537	0.512	0.534	0.537
HOG + MFV	0.432	0.433	0.459	0.432	0.489	0.462	0.503	0.489
All	0.259	0.247	0.330	0.259	0.313	0.277	0.341	0.313

Table 3. Classification results using concatenated MFV representations.

Model	Acc.	F1	Prec.	Rec.
SVM	0.651	0.650	0.655	0.650
MLP	**0.653**	**0.651**	**0.653**	**0.653**
RF	0.631	0.622	0.633	0.631
KNN	0.558	0.551	0.569	0.558

Table 4. Best voting results.

Input	Combination	Acc.	F1	Prec.	Rec.
Preproc.	RF + MLP, Soft	0.637	0.633	0.637	0.637
Raw	RF + SVM + MLP	0.551	0.546	0.552	0.551
Concat. MFV	RF + MLP, Soft	**0.666**	**0.664**	**0.667**	**0.666**

However, slightly better performance was obtained when the models were trained using raw centrally cropped images, as shown in Table 5. In this setting, accuracy ranged from 80.5% to 83.3%. These results suggest that the preprocessing may remove subtle visual information that deep neural networks are able to exploit during feature learning.

Table 5. Results from the CNNs.

	With preprocessing				Raw cut images			
Modelo	Acc.	F1	Prec.	Rec.	Acc.	F1	Prec.	Rec.
AlexNet	0.784	0.783	0.785	0.784	0.805	0.804	0.806	0.805
EfficientNet	0.783	0.783	0.791	0.783	0.810	0.810	0.813	0.810
MobileNet	**0.815**	**0.814**	**0.815**	**0.815**	**0.833**	**0.833**	**0.834**	**0.833**
ResNet	**0.814**	**0.815**	**0.816**	**0.814**	**0.832**	**0.832**	**0.833**	**0.832**

5.3. Hybrid approach

To further investigate the complementarity between handcrafted and deep representations, we evaluated a hybrid ensemble combining both approaches. In this configuration, Random Forest and MLP were used for the handcrafted features (based on the MFV representation), while ResNet50 and MobileNetV2 represented the CNN-based models. Different voting weights were explored through grid search on the training set to determine the best soft voting configuration.

The hybrid model significantly improved the classification performance, increasing accuracy from 0.833 (best individual CNN) to 0.929, as shown in Table 6. Interestingly, the optimal weight configuration (Table 7) assigns the largest weight to the Random Forest classifier ($w = 0.40$). This may indicate that, although MLP achieved the best performance among handcrafted models individually, its predictions may be more correlated with those of the CNNs, contributing less additional information to the ensemble.

In terms of computational cost, we found that, using ResNet50's average inference time as a baseline (12.2ms), the inference time of the MobileNet was faster, at 0.64x of the baseline. Furthermore, the feature extraction and preprocessing steps measured at 2.77x and 4.92x, respectively, being the most costly. The inference by RF and MLP were the fastest, at ≤ 0.002x.

Despite these costs, the preprocessing, feature extraction and inference steps are well-suited for a pipelined architecture. Implementing such a pipeline would significantly increase throughput for large-scale datasets by overlapping these stages, though this optimization remains outside the scope of the current work

5.4. Discussion

Even though comparing our results with previous studies suggests potential improvements, direct comparisons are not entirely straightforward. The GZ2 dataset has grown over time, and different studies adopt distinct criteria for defining galaxy classes, which may affect the number of categories and the class distribution used in the experiments.

In this work, we adopted the same seven-class classification scheme proposed by [Kalvankar et al. 2020], which extends the five-class taxonomy used

Table 6. Results from the hybrid approach.

Class	Precision	Recall	F1
0	0.953	0.944	0.949
1	0.912	0.915	0.914
2	0.872	0.931	0.900
3	0.921	0.900	0.911
4	0.973	0.962	0.967
5	0.925	0.904	0.914
6	0.948	0.944	0.946
Accuracy			0.929

Table 7. Weights used in the optimal combination of the hybrid approach.

	MLP	RF	ResNet	MobileNet
Weight	0.10	0.40	0.30	0.20

by [Zhu et al. 2019], thus making the classification task more challenging. Table 8 summarizes the best classification rates obteined on those works, as to briefly show the performances obtained in them.

Although our overall accuracy is approximately two percentage points lower than that reported by [Zhu et al. 2019], our results are competitive when considering the more challenging seven-class setting. In addition, the class-wise performance in our model is relatively consistent across all classes. In particular, the lowest F1-score obtained in our experiments is 0.900 for class 2 (cigar-shaped smooth galaxies), whereas [Zhu et al. 2019] report an F1-score of 0.647 and [Kalvankar et al. 2020] report 0.670 for the same class. This suggests that the hybrid approach explored in this work may provide a more balanced performance across classes, particularly benefiting categories that are typically more difficult to classify.

These results are consistent with the hypothesis explored in this work that handcrafted descriptors and deep representations capture complementary information, which can be effectively exploited through hybrid ensemble strategies.

Table 8. Classification rates presented in this work vs. literature.

Work	# Classes	Acc.	F1	Prec.	Recall
[Zhu et al. 2019]	5	0.946	0.951	0.951	0.952
[Kalvankar et al. 2020]	7	0.937	0.885	0.888	0.890
This work	7	0.929	0.929	0.929	0.929

6. Conclusion

In this work, we evaluated handcrafted and deep learning approaches for galaxy morphology classification using a balanced subset of the GZ2 dataset. Our experiments investigated the impact of a multi-step preprocessing pipeline, the effectiveness of handcrafted descriptors combined with classical machine learning algorithms, and the performance of transfer-learning CNN models.

The results show that deep learning models significantly outperform traditional machine learning methods. In particular, MobileNetV2 and ResNet50 achieved an F1-score of 0.833 when trained on raw centrally cropped images. Interestingly, although the preprocessing pipeline improved the performance of classical machine learning models, it slightly reduced the performance of CNN-based models, suggesting that deep architectures are able to exploit subtle visual patterns that may be attenuated during manual preprocessing.

The most important finding of this study is that combining handcrafted and deep representations leads to further improvements. By adopting a weighted soft-voting ensemble, the hybrid approach achieved an average F1-score of 0.929. Moreover, the results were relatively consistent across all seven classes, particularly improving the classification of the “cigar-shaped smooth” galaxies, which are typically more difficult to identify.

Beyond the specific application to galaxy morphology classification, this study reflects a broader transformation occurring in data-intensive scientific domains. Modern astronomy increasingly relies on scalable computational pipelines capable of processing massive image repositories generated by digital observatories. The development of efficient machine learning workflows for large-scale image analysis therefore contributes not only to astrophysical research, but also to the broader ecosystem of data-driven scientific discovery.

Future work may explore deeper architectures, alternative ensemble strategies among CNN models, different class balancing techniques, interpretability analysis along XAI (eXplainable AI), and a pipelined architecture for enhanced computational efficiency. Furthermore, tests on different, non-balanced datasets can be explored as well. These directions may further strengthen the role of intelligent systems in large-scale scientific image processing.

Acknowledgement

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001 and by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

References

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR’05)*, volume 1, pages 886–893. Ieee.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Fortson, L., Masters, K., Nichol, R., Borne, K., Edmondson, E., Lintott, C., Raddick, J., Schawinski, K., and Wallin, J. (2011). Galaxy zoo: Morphological classification and citizen science.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Hubble, E. P. (1926). Extragalactic nebulae. *Astrophysical Journal, 64, 321-369 (1926)*, 64.

Kalvankar, S., Pandit, H., and Parwate, P. (2020). Galaxy morphology classification using efficientnet architectures. *arXiv preprint arXiv:2008.13611*.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Ntwaetsile, K. and Geach, J. E. (2021). Rapid sorting of radio galaxy morphology using haralick features. *Monthly Notices of the Royal Astronomical Society*, 502(3):3417–3425.

Ojansivu, V. and Heikkilä, J. (2008). Blur insensitive texture classification using local phase quantization. In *International conference on image and signal processing*, pages 236–243. Springer.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520.

Shamir, L. (2009). Automatic morphological classification of galaxy images. *Monthly Notices of the Royal Astronomical Society*, 399(3):1367–1372.

Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR.

Trickz, J. (2020). Galaxy Zoo 2: Images. `https://www.kaggle.com/datasets/jaimetrickz/galaxy-zoo-2-images`.

Willett, K. W., Lintott, C. J., Bamford, S. P., Masters, K. L., Simmons, B. D., Casteels, K. R. V., Edmondson, E. M., Fortson, L. F., Kaviraj, S., Keel, W. C., Melvin, T., Nichol, R. C., Raddick, M. J., Schawinski, K., Simpson, R. J., Skibba, R. A., Smith, A. M., and Thomas, D. (2013). 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.

Zhu, X.-P., Dai, J.-M., Bian, C.-J., Chen, Y., Chen, S., and Hu, C. (2019). Galaxy morphology classification with deep convolutional neural networks. *Astrophysics and Space Science*, 364(4).