Deep Learning for Classification of Low Surface Brightness Galaxies in Dark Energy Survey

Bilguun Batbayar[1]
Department of Astronomy and Astrophysics
Kavli Institute for Cosmological Physics
University of Chicago
bilguun@uchicago.edu

Alex Drlica-Wagner
Department of Astronomy and Astrophysics
Kavli Institute for Cosmological Physics
University of Chicago
kadrlica@uchicago.edu

Kai Herron
Department of Physics and Astronomy
Dartmouth College
kai.r.herron.gr@dartmouth.edu

Burcin Mutlu-Pakdil
Department of Physics and Astronomy
Dartmouth College
Burcin.Mutlu-Pakdil@dartmouth.edu

Abstract

Low Surface Brightness Galaxies (LSBGs) are faint, diffuse systems that are difficult to identify and are often confused with imaging artifacts in wide-field surveys. In this work, we apply convolutional neural networks (CNNs) to classify LSBGs in the full six-year Dark Energy Survey (DES Y6) dataset, extending earlier analyses on the Year 3 data. Training on the $\sim$40,000 labeled Y6 objects used in prior work, our CNN achieves 90.7% accuracy, surpassing traditional feature-based methods. Applied to the full Y6 sample of $\sim$70,000 labeled-objects, the network reliably identifies artifacts, but struggles with ambiguous human-labeled LSBG cases. Targeted augmentation reduces the fraction of LSBGs classified as artifacts from 27.4% to 18.5%, bringing CNN predictions into closer alignment with human labels. We also compare two classifiers that highlight a trade-off: one favors completeness by retaining more LSBG candidates (at the risk of inaccuracy), while the other favors purity by excluding ambiguous cases. Overall, CNNs classify LSBGs with high efficiency and accuracy while also uncovering potential human mislabels. With improved training data and stronger architectures, CNN-based approaches will be indispensable for understanding the low-surface-brightness universe in future large-scale surveys.

1 Introduction

1.1 Low surface brightness universe and Identification of LSBGs

Low Surface Brightness Galaxies (LSBGs) are diffuse stellar systems whose average luminosity per unit area falls below that of the night sky, often by more than a magnitude. Their intrinsic faintness makes them difficult to detect in traditional surveys, and only in the past few decades have they been systematically studied [Impey and Bothun, 1997]. LSBGs are also among the most dark-matter-dominated systems known, making them valuable laboratories for testing models of dark matter and alternative theories of gravity. Studying these galaxies expands the census of galaxy populations while providing constraints on galaxy formation, feedback processes, and the distribution of dark matter in the universe [de Blok and McGaugh, 1997].With modern surveys and deep-imaging techniques, the number of known LSBGs continues to grow, offering new opportunities for refining cosmological models.

Machine Learning and the Physical Sciences Workshop, NeurIPS 2025.

The detection of LSBGs presents a major challenge in observational astronomy. Owing to their diffuse profiles, LSBGs can be confused with a variety of imaging artifacts, including scattered light from bright stars, cosmic rays, Galactic cirrus, and residual background fluctuations [Koda et al., 2015]. Historically, the separation of real LSBGs from artifacts has relied heavily on visual inspection. While effective in limited cases, this approach is subjective and ultimately impractical given the data volumes produced by current surveys. These limitations have motivated the development of automated classification methods. In particular, deep learning approaches based on convolutional neural networks (CNNs) have achieved notable success in astronomical image analysis. For instance, CNNs applied to the CFHTLS-Wide Survey detected faint tidal features in galaxies, outperforming conventional methods when combined with regularization and data augmentation techniques [Walmsley et al., 2019].

1.2 CNN-based classification of DES LSBGs

Tanoglidis et al. introduced *DeepShadows*, a CNN trained on Dark Energy Survey (DES) Year 3 (Y3) data to separate LSBGs from artifacts [Tanoglidis et al., 2020]. Using a large sample of visually inspected galaxies and artifacts, the network achieved 92% accuracy, outperforming traditional ML methods. Applied to Hyper Suprime-Cam (HSC) data, it reached 82.1% without retraining and 87.6% with fine-tuning, demonstrating both strong performance within DES and adaptability across surveys.

Relative to previous work, our study is the first to systematically evaluate CNN performance on the full DES Year 6 (Y6) dataset, which features deeper imaging, improved photometric calibration, and a more diverse artifact population. These properties both increase sensitivity to faint galaxies and create a more challenging, realistic testbed for assessing CNN classifier robustness.

2 Data and Methods

2.1 The Dark Energy Survey

The Dark Energy Survey (DES) is an optical and near-infrared imaging survey aimed at probing dark energy through large-scale structure, galaxy clusters, and weak lensing. Using the 570-megapixel Dark Energy Camera (DECam) on the 4-m Blanco Telescope in Chile, DES imaged $\sim$5,000 deg^2 of the southern sky in five bands (*grizY*) between 2013 and 2019 [Abbott and the DES Collaboration, 2018]. The Year 6 (Y6) Gold catalog is the deepest and most uniform DES dataset to date, containing $\sim$669 million objects over nearly the full footprint. It reaches a depth of $i_{AB} \approx 23.4$ mag for extended sources at $\mathrm{S/N} \approx 10$ and delivers photometric uniformity better than 2 mmag [Bechtol and Collaboration, 2025]. Compared to Year 3 (Y3), Y6 increases sensitivity to LSBGs while also introducing more complex noise properties and artifact diversity.

2.2 LSBG catalog

The LSBG catalog was constructed in four stages:

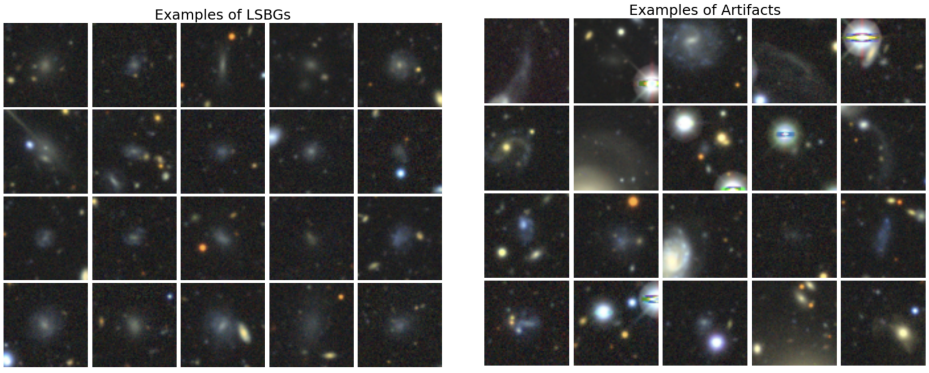


Figure 1: Representative cutouts from DES Y6. Left: low surface brightness galaxies (LSBGs); Right: artifacts.

1. Apply selection cuts using SOURCEEXTRACTOR parameters [Bertin and Arnouts, 1996].
2. Use a random forest classifier on these features, identifying ∼80,000 candidates.
3. Perform visual inspection to reject false positives, leaving ∼50,000 LSBGs and ∼30,000 artifacts.
4. Fit single Sérsic profiles with GALFIT [Peng et al., 2002], yielding ∼40,000 galaxies with reliable light profiles.

The Y6 sample thus contains ∼40,000 modeled LSBGs and ∼30,000 artifacts. For comparison, the Y3 analysis used ∼40,000 visually labeled objects ($20,000$ LSBGs and $20,000$ artifacts) [Tanoglidis et al., 2020]. We matched those Y3 objects to their Y6 counterparts; nearly all were recovered in the Y6 sample (Fig. 1).

2.3 CNN Architecture

We adopt a convolutional neural network (CNN) for binary classification of 64×64 RGB cutouts (LSBG vs. artifact), following Tanoglidis et al. [2020]. The model has three convolutional blocks with increasing filters (16, 32, 64), each with ReLU activation, 3×3 kernels, max pooling, dropout, and L_2 regularization. A dense layer (1024 units) precedes the final sigmoid output. Batch normalization is used throughout, and accuracy is the primary evaluation metric. Hyperparameters match those in Table 1. We chose this architecture because it closely follows the DeepShadows model and enables a direct comparison between Y3 and Y6, and preliminary tests with deeper CNN architectures did not yield clear improvements in validation accuracy.

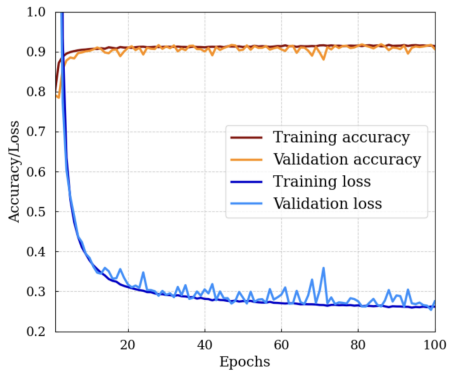
Table 1: CNN architecture used in this work.

Stage	**Layers**	**Output**
Input	$64 \times 64 \times 3$ RGB	$64 \times 64 \times 3$
Block 1	Conv2D (16, 3×3), ReLU, L_2=0.13, MaxPool 2×2, Dropout 0.4	$32 \times 32 \times 16$
Block 2	Conv2D (32, 3×3), ReLU, L_2=0.13, MaxPool 2×2, Dropout 0.4	$16 \times 16 \times 32$
Block 3	Conv2D (64, 3×3), ReLU, L_2=0.13, MaxPool 2×2, Dropout 0.4	$8 \times 8 \times 64$
Flatten	—	4096
Dense	Dense(1024), ReLU, L_2=0.12	1024
Output	Dense(1), Sigmoid	1

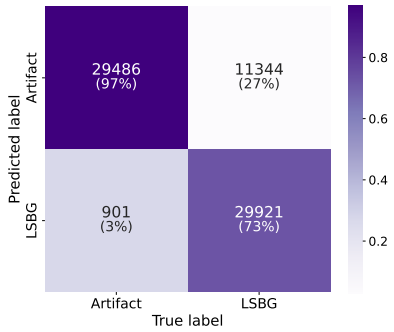
3 Classification results

3.1 Main findings

We first test whether the Year 3 results can be reproduced on the same objects in Year 6. RGB cutouts of ∼40K "Y3 objects in Y6" were split into training, validation, and test sets (30K, 5K, and 5K



(a) Training and validation accuracy/loss.



(b) Confusion matrix on ∼40K LSBGs and ∼30K artifacts.

Figure 2: Training history (left) and CNN confusion matrix (right).

objects, respectively) and used to train the CNN described earlier. As shown in Fig. 2a, the training history indicates stable convergence without overfitting. The model achieved 90.7% accuracy on the test set. For comparison, SVM and Random Forest models trained on SOURCEEXTRACTOR features reached 82.9% and 80.6% accuracy.

We next evaluate the CNN on the full DES Y6 dataset of ∼50K galaxies and 30K artifacts. After removing ∼10K galaxies with poor GALFIT fits or visual flags, the sample includes ∼40K LSBGs and ∼30K artifacts. The classification results are shown in Fig. 2b. The network reliably identifies artifacts, but 27% of human-labeled LSBGs are classified as artifacts. Re-inspection of low-confidence cases shows that many of these are in fact genuine LSBGs. This indicates that the network mainly struggles with ambiguous LSBGs, often classifying them as artifacts.

These ambiguous cases typically have very low central surface brightness, irregular morphologies, and diffuse light profiles that approach the local background. They may also be contaminated by light from nearby galaxies, making them difficult to distinguish. Improving labels for these borderline objects is therefore essential for further progress.

To address this, we explore augmenting the training dataset. A set of 380 cases that the network classified as artifacts but were confirmed as genuine LSBGs were expanded through flips (Fig. 3), yielding ∼1.5K augmented images. Retraining with this data reduced the fraction of LSBGs classified as artifacts from 27.4% to 18.5%. These results demonstrate that targeted augmentation improves agreement between CNN predictions and human labels.

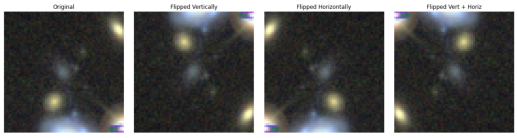


Figure 3: Example of data augmentation: original, vertical flip, horizontal flip, combined flip.

Then we compared two classifiers: **DeepSkyNet-Y6**, trained on ∼70K labeled Y6 objects, and **DeepSkyNet-Y3**, trained on 40K "Y3 objects in Y6" with augmentation. Both were evaluated on the same ∼20K test objects for comparison. As shown in Fig. 4, both classifiers performed well in identifying human-labeled artifacts. For human-labeled LSBGs, DeepSkyNet-Y6 retained more candidates, classifying 93% as LSBGs, while DeepSkyNet-Y3 retained only 76%. We also observed particularly interesting cases where both classifiers disagreed with the human labels—for example, objects labeled as artifacts but identified by both classifiers as LSBGs. Such cases likely point to mislabels in the original dataset.

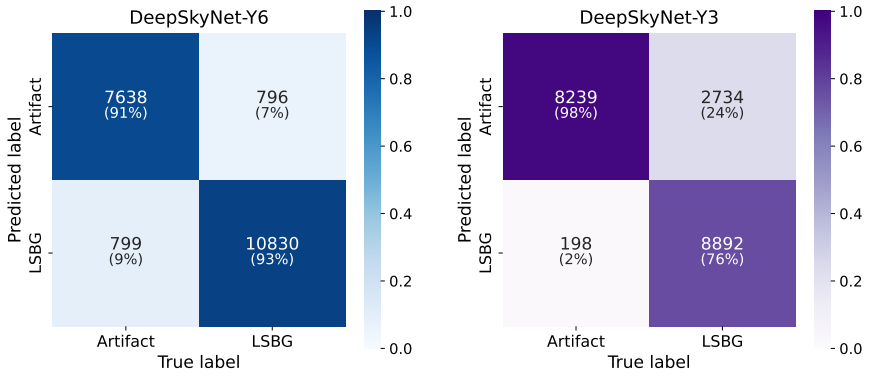


Figure 4: Confusion matrices summarizing the classification of ∼20K test objects for DeepSkyNet-Y6 (left, Blues) and DeepSkyNet-Y3 (right, Purples).

4 Discussion and Conclusions

In this work, we applied deep learning to classify low-surface-brightness galaxies (LSBGs) and artifacts in the full six-year DES dataset. Our results confirm and extend earlier Year 3 findings: CNNs

achieve high accuracy, outperform traditional machine-learning pipelines, and expose inconsistencies in human labeling. Performance depends critically on the quality of the training data, but targeted augmentation and retraining already yield predictions that align more closely with human judgments. This demonstrates that the network's behavior can be shaped through careful training, allowing it to classify in ways that better mirror human judgement, as seen in the comparison between the two classifiers.

The main uncertainties in our analysis arise from the quality of the visually inspected labels. Looking ahead, the most immediate step is refining the DES Year 6 labels by correcting clear mistakes and flagging ambiguous cases. A finalized catalog will enable these networks to be applied across surveys to test generalizability and reduce reliance on manual inspection. At the same time, exploring alternative CNN architectures and systematic hyperparameter optimization, ideally with parallelized training, offers avenues for further gains.

In current DES workflows, CNNs are not intended to replace human inspection entirely but to substantially reduce the number of candidates requiring manual review. Even in the presence of imperfect labels, the classifier meaningfully accelerates LSBG catalog construction by filtering out clear artifacts and highlighting cases worthy of additional inspection.

Looking ahead, our findings have direct implications for upcoming surveys such as LSST and Euclid, which will produce vastly larger samples of faint, diffuse galaxies. The sensitivity of CNN performance to training-set construction suggests high-quality labeled training data. Overall, ML/AI methods have the immense potential to drive discovery in large-scale scientific surveys, and with better data and models, CNN-based approaches will undoubtedly advance the study of low-surface-brightness galaxies.

References

T. Abbott and the DES Collaboration. The dark energy survey: more than dark energy—an overview. *Monthly Notices of the Royal Astronomical Society*, 460:1270–1299, 2018. (Commonly cited DES overview; use the version appropriate to your context).

K. Bechtol and DES Collaboration. Dark energy survey year 6 results: Photometric data set for cosmology (y6 gold). *arXiv preprint*, 2025.

E. Bertin and S. Arnouts. Sextractor: Software for source extraction. *Astronomy & Astrophysics Supplement Series*, 117:393–404, 1996. doi: 10.1051/aas:1996164.

W. J. G. de Blok and S. S. McGaugh. The dark and baryonic matter content of low surface brightness disk galaxies. *Monthly Notices of the Royal Astronomical Society*, 290:533–552, 1997. doi: 10.1093/mnras/290.3.533.

C. D. Impey and G. D. Bothun. Low surface brightness galaxies. *Annual Review of Astronomy and Astrophysics*, 35:267–307, 1997. doi: 10.1146/annurev.astro.35.1.267.

J. Koda et al. Approximately a thousand ultra-diffuse galaxies in the coma cluster. *Astrophysical Journal Letters*, 807:L2, 2015. doi: 10.1088/2041-8205/807/1/L2.

C. Y. Peng, L. C. Ho, C. D. Impey, and H.-W. Rix. Detailed structural decomposition of galaxy images. *Astronomical Journal*, 124:266–293, 2002. doi: 10.1086/340952.

D. Tanoglidis, A. Ćiprijanović, and A. Drlica-Wagner. Deepshadows: Separating low surface brightness galaxies from artifacts using deep learning. *Astronomy and Computing*, 35:100469, 2020. doi: 10.1016/j.ascom.2021.100469.

M. Walmsley et al. Identification of low surface brightness tidal features in galaxies using convolutional neural networks. *Monthly Notices of the Royal Astronomical Society*, 491:1554–1574, 2019.