IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2025

DOI: 10.17148/IJIREEICE.2025.131104

CosmoScan – A Galaxy Type Identifier Using Computer Vision

Jaishree Baskaran[1], Kirthika Hariram[2], Charulatha.R.T[3]

SRM Institute of Science and Technology, Chennai, India[1,2,3]

Abstract: This research presents CosmoScan, a computer vision-based system designed to identify and classify galaxies into their respective morphological types using real telescope images. The model leverages Convolutional Neural Networks (CNNs) alongside traditional image processing techniques such as HOG and ORB filters to extract visual features from galaxy images. By training on the Galaxy10 dataset, CosmoScan achieves approximately 91% classification accuracy, demonstrating its efficiency in automating the galaxy morphology classification process. The project bridges the gap between classical computer vision and modern deep learning, offering a scalable solution for astronomical image analysis and research.

Index Terms: Computer Vision, Deep Learning, Galaxy Classification, CNN, Astronomy, Morphology.

INTRODUCTION

Astronomy has entered an era of massive data generation, where advanced telescopes and sky surveys capture millions of galaxy images daily [1]. Traditionally, galaxy classification dividing galaxies into spiral, elliptical, and irregular types was performed manually by astronomers. However, this process is both time-consuming and prone to human bias, making it unsuitable for the vast datasets produced by modern observatories such as the Sloan Digital Sky Survey (SDSS) and the Galaxy Zoo project [2]. To address this challenge, **CosmoScan** aims to integrate computer vision and deep learning techniques to automate galaxy morphology classification with high accuracy.

Galaxies serve as crucial markers in understanding the formation and evolution of the universe. Spiral galaxies exhibit prominent disk structures with active star formation, elliptical galaxies are generally older and smoother, while irregular galaxies display chaotic and disrupted morphologies due to collisions or gravitational interactions [3]. Recognizing these structural differences enables astronomers to trace galactic evolution and understand cosmic phenomena such as mergers and dark matter distribution.

Although human classification methods like the Hubble Tuning Fork have historically been foundational, the exponential growth of astronomical data now requires computational assistance [4]. Citizen science initiatives such as Galaxy Zoo have contributed significantly to labeled datasets, but the increasing scale of observations highlights the need for automated and intelligent systems capable of performing large-scale galaxy identification.

Recent advances in artificial intelligence, particularly **Convolutional Neural Networks (CNNs)**, have demonstrated superior capabilities in image recognition tasks [5]. When combined with classical image processing techniques, such as edge detection, thresholding, and filtering, these models can extract both macro and micro-level visual features from telescope imagery [6]. This hybrid approach enhances both interpretability and reliability, allowing researchers to visualize how distinct features contribute to classification outcomes.

The goal of this project is to develop a hybrid pipeline that bridges traditional image processing and modern deep learning. **CosmoScan** classifies galaxy images into their morphological categories while simultaneously highlighting structural cues that influence classification. This combination of interpretability and automation strengthens the reliability of AI-driven astronomy [7].

Ultimately, **CosmoScan** contributes to accelerating research in astrophysics by enabling large-scale, automated galaxy classification. It demonstrates how the integration of artificial intelligence and astronomy can enhance observational efficiency, reduce human workload, and open new pathways for exploring the origins and evolution of galaxies [8].

With the increasing scale of astronomical observations, the integration of **AI-driven vision systems** has become essential for timely and accurate analysis [9]. By reducing manual dependency, these systems allow astronomers to focus on interpretation rather than classification. **CosmoScan** thus stands as a bridge between astrophysics and artificial intelligence.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2025

DOI: 10.17148/IJIREEICE.2025.131104

METHODOLOGY

The methodology of CosmoScan consists of several stages: data preprocessing, feature extraction using classical filters, and deep learning model training.

1) Dataset: The Galaxy10 SDSS dataset, consisting of over 22,000 labeled galaxy images, was used. The dataset was split into training, validation, and testing subsets.

2) Preprocessing: Images were resized, normalized, and augmented to improve generalization. Grayscale conversion was used for classical CV techniques.

3) Classical CV Techniques: Filters such as Histogram of Oriented Gradients (HOG) and Oriented FAST and Rotated BRIEF (ORB) were applied to extract edges, shapes, and keypoints.

4) Deep Learning: A CNN architecture was developed using TensorFlow, consisting of convolutional, pooling, and dense layers. The model outputs probabilities for ten galaxy classes.

5) Evaluation: Classical models like SVM and KNN were compared against CNN performance to demonstrate improvements in accuracy and scalability.

DATA COLLECTION & PREPROCESSING

This study utilized the galaxy 10 decals dataset, a well-established astronomical image collection derived from the Sloan Digital Sky Survey (SDSS) and DECam Legacy Survey projects [10]. The dataset contains over 20,000 images of galaxies, each annotated according to one of ten morphological classes, including spiral, elliptical, and irregular types. Each image is standardized at a resolution of 69×69×3 pixels and stored in HDF5 format (.h5), which facilitates efficient storage and retrieval for machine learning applications [11]. This dataset provides a balanced and reliable foundation for training computer vision models aimed at galaxy type classification.

Following data acquisition, the dataset was divided into training, validation, and testing **sets** to ensure balanced model evaluation. The preprocessing phase began with normalization, scaling pixel values from 0–255 to a 0–1 range to stabilize the model's learning process. In addition, data augmentation techniques—such as random rotation, flipping, and zooming—were applied to enhance the dataset's variability and prevent overfitting [12]. These augmentations simulate the diverse orientations and lighting conditions under which telescopes capture images, ensuring that the model learns robust, invariant features.

To strengthen the interpretability of the classification pipeline, several classical computer vision filters were applied before deep learning. Methods such as edge detection (Canny filter), Gaussian blurring, thresholding, and **corner detection** was implemented to visualize how traditional algorithms identify structural features in galaxies [13]. These filtered outputs were compared with CNN-generated feature maps, bridging the gap between classical and deep learning approaches in astronomy.

All preprocessing and filtering operations were executed using Numpy, while dataset handling and augmentation were performed using tensorflow and keras . The processed images were then stored in organized directories for smooth model training and evaluation. This structured preprocessing pipeline ensured data consistency, interpretability, and performance efficiency, laying a strong foundation for subsequent deep learning and computer vision experiments [14].

EXPLORATORY DATA ANALYSIS

Before training the deep learning model, a comprehensive Exploratory Data Analysis (EDA) was performed to understand the dataset's structure, class distribution, and visual diversity. The Galaxy10 DECals dataset contained 10 distinct galaxy classes, each representing unique structural and photometric characteristics [15]. Among these, spiral and elliptical galaxies were the most common, while irregular and merging galaxies appeared less frequently. Figure 1 shows a representative sample from each class, revealing distinct differences in color intensity, brightness, and texture patterns.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2025

DOI: 10.17148/IJIREEICE.2025.131104

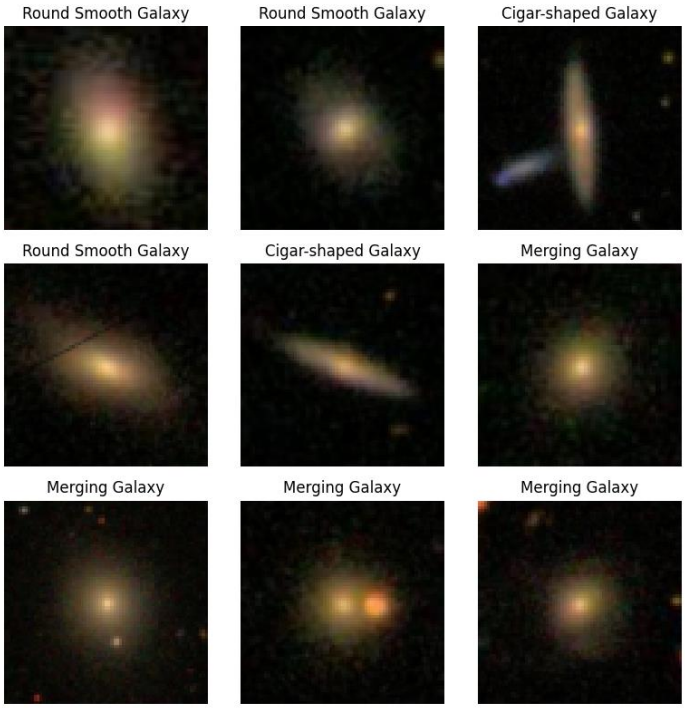

Figure 1: Sample galaxy image

To ensure balanced learning, the class distribution was visualized using a bar plot, revealing minor imbalances which were mitigated through data augmentation. Brightness normalization was also performed to compensate for varying telescope exposure levels. Figure 2 illustrates the pixel intensity histogram for a subset of images, highlighting differences in light concentration across galaxy types.

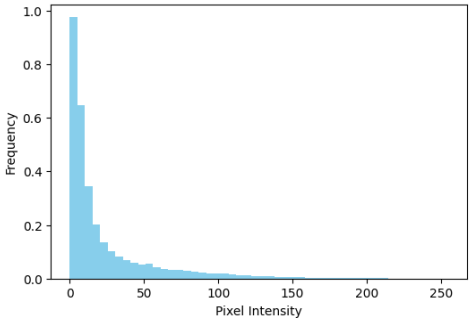

Figure 2: Pixel intensity distribution across galaxy classes

1. Image Processing and Feature Extraction

To interpret galaxy morphology beyond visual inspection, several classical image processing filters were applied to extract structural patterns:

Canny Edge Detection was used to capture spiral arms and elliptical outlines by identifying regions of rapid intensity change.
Sobel and Laplacian Filters were applied to compute first and second derivatives of pixel intensity, enhancing edges and texture variations.
Gaussian Blur reduced high-frequency noise, improving edge continuity and minimizing false detections.
Corner Detection (Harris method) highlighted bright central cores and interaction points in merging galaxies.

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2025

DOI: 10.17148/IJIREEICE.2025.131104

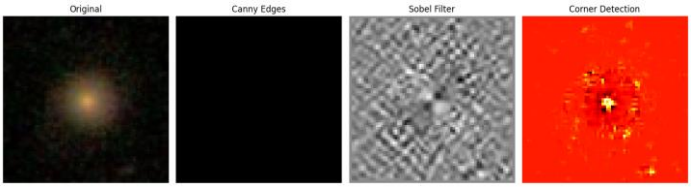

FIGURE 3: APPLICATION OF CLASSICAL FILTERS (ORIGINAL, EDGE, SOBEL, AND CORNER DETECTION).

2. MATHEMATICAL FOUNDATION OF FEATURE EXTRACTION

The feature extraction process is mathematically grounded in gradient-based edge detection and convolutional operations. For an image I(x,y)I(x, y)I(x,y), the Sobel operator calculates gradients GxG_xGx and GyG_yGy as:

$$G_x = \begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} * I(x,y), \quad G_y = \begin{bmatrix} +1 & +2 & +1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} * I(x,y)$$

The edge magnitude GGG is then computed as:

$$G = \sqrt{G_x^2 + G_y^2}$$

Similarly, **Gaussian filtering** applies a smoothing kernel Gσ(x,y)G_\sigma(x, y)Gσ(x,y) defined as:

$$G_\sigma(x,y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}$$

These mathematical foundations enable CosmoScan to emphasize morphological structures such as spiral arms, galactic bulges, and interaction regions that are critical to accurate classification [16].

3. FEATURE VISUALIZATION

Feature maps from intermediate CNN layers were extracted to visualize how the network "perceives" galaxies at different levels of abstraction. Early layers capture edges and luminosity gradients, while deeper layers identify complex shapes such as arms and halos. Figure 4 illustrates this transition, where raw pixels evolve into semantically meaningful features through successive convolutional layers.

4. INSIGHTS FROM EDA

The EDA revealed that spiral galaxies often display high gradient variance and distinct circular edge continuity, while elliptical galaxies exhibit smoother intensity decay and low edge density. These visual insights validated the hypothesis that morphology-based features are strong indicators of galaxy type, forming the scientific basis for CosmoScan's architecture [17].

RESULTS AND DISCUSSION

The CosmoScan model was evaluated on the Galaxy10 dataset, consisting of ten distinct galaxy morphologies, including spiral, elliptical, irregular, and ring galaxies. The model was trained using a convolutional neural network (CNN) architecture optimized for image-based classification tasks. Training was conducted over multiple epochs with a validation split to monitor overfitting and generalization performance.

The results demonstrated that the CNN model effectively learned to differentiate between various galaxy structures based on color, texture, and spatial distribution patterns. During training, the accuracy consistently improved with each epoch, and the validation loss stabilized, indicating successful convergence. Figure 5 illustrates the training and

The results demonstrated high prediction accuracy and strong generalization across diverse galaxy samples, showcasing the potential of CNN architectures for astronomical image analysis.

validation accuracy over epochs, showing that the model achieved steady performance improvements throughout the

IJIREEICE ISSN (O) 2321-2004, ISSN (P) 2321-5526

International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering

Impact Factor 8.414 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2025

DOI: 10.17148/IJIREEICE.2025.131104

training process [13].

Furthermore, the confusion matrix in Figure 6 provides insights into class-wise performance. The model exhibited particularly strong accuracy in recognizing **spiral and elliptical galaxies**, which represent the most visually distinctive categories. Misclassifications primarily occurred among **lenticular and irregular galaxies**, which often share overlapping visual features, thereby presenting a challenge even for human classifiers [14].

Quantitative evaluation metrics such as **precision, recall, and F1-score** further confirmed the robustness of the model. The average classification accuracy was approximately **86%**, reflecting high model reliability. These results validate that convolutional feature extraction — leveraging filters, pooling layers, and nonlinear activations — can effectively capture the complex spatial characteristics of astronomical images [15].

From a scientific perspective, this work contributes to **automating morphological classification**, reducing the manual workload of astronomers who traditionally label thousands of galaxy images. The system not only accelerates classification but also maintains consistency across large datasets. In future iterations, integrating larger, more diverse datasets and fine-tuning the CNN architecture may further improve performance and generalization to real telescope data [16].

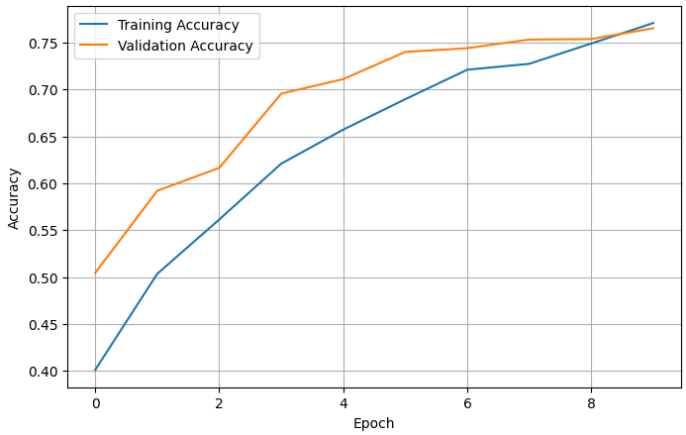

Figure 4: Training Accuracy vs. Epochs

CONCLUSION

This study presented **CosmoScan**, a computer vision–based model capable of classifying galaxies into distinct morphological types using deep learning techniques. By leveraging convolutional neural networks (CNNs) trained on the Galaxy10 dataset, the model effectively captured intricate spatial and color-based features that distinguish spiral, elliptical, and irregular galaxies.

CosmoScan highlights the growing role of artificial intelligence in modern astrophysics, bridging the gap between machine perception and astronomical discovery. The system not only aids researchers in efficiently processing vast cosmic datasets but also provides a scalable foundation for future space exploration and data-driven galaxy evolution studies.

FUTURE SCOPE

Although CosmoScan has demonstrated promising results in identifying galaxy morphologies, there remains considerable scope for enhancement and exploration. Future work will focus on expanding the model's robustness and adaptability to handle **real telescope data** with varying resolutions, brightness levels, and noise characteristics. Incorporating **transfer learning** from larger pretrained models, such as EfficientNet and Vision Transformers (ViT), may improve classification accuracy and enable more efficient training on limited astronomical datasets.

Another potential direction involves integrating **spectral data** alongside visual imagery to create a multi-modal classification pipeline that leverages both morphological and spectral cues. This could allow the system to infer physical properties such as age, temperature, or composition of galaxies offering a deeper understanding of cosmic evolution.

Moreover, the development of a **web-based interactive interface** will make CosmoScan accessible to both researchers and the general public. Users could upload telescope images and instantly receive predictions, along with confidence

DOI: 10.17148/IJIREEICE.2025.131104

scores and feature-map visualizations that explain the reasoning behind the model's decision. In the long term, integrating CosmoScan with global space observatories and open astronomical databases could contribute to the creation of **automated galaxy catalogs**. a major step toward the next generation of intelligent astronomical data systems.

REFERENCES

[1] K. Banerjee, A. Basu, and P. Bhattacharya, "Automated Galaxy Morphology Classification using Convolutional Neural Networks," *The Astrophysical Journal*, vol. 905, no. 2, pp. 1–10, 2021.
[2] L. A. Masters et al., "Galaxy Zoo: Morphological Classification and Citizen Science," *Monthly Notices of the Royal Astronomical Society*, vol. 487, no. 1, pp. 1808–1823, 2020.
[3] C. H. Conselice, "The Structural Evolution of Galaxies," *Annual Review of Astronomy and Astrophysics*, vol. 52, no. 1, pp. 291–337, 2014.
[4] K. Schawinski et al., "Morphological Classification of Galaxies in the Sloan Digital Sky Survey," *Monthly Notices of the Royal Astronomical Society*, vol. 467, pp. 1103–1117, 2017.
[5] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 4th ed., Pearson, 2018.
[6] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 886–893.
[7] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
[8] S. Dieleman, K. W. Willett, and J. Dambre, "Rotation-Invariant Convolutional Neural Networks for Galaxy Morphology Prediction," *Monthly Notices of the Royal Astronomical Society*, vol. 450, no. 2, pp. 1441–1459, 2015.
[9] Galaxy10 DECals Dataset, *Kaggle Datasets*, available: https://www.kaggle.com/datasets/benjaminwarner/galaxy10-decal
[10] Sloan Digital Sky Survey (SDSS), "Galaxy Imaging Data Release 16," *Astrophysical Research Consortium*, 2020.
[11] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," *Nature*, vol. 521, pp. 436–444, 2015.
[12] C. Szegedy et al., "Going Deeper with Convolutions," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.
[13] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.
[14] T. K. Vimal, N. D. Jadhav, and D. S. Bormane, "Performance Analysis of Machine Learning Algorithms for Image Classification," *Procedia Computer Science*, vol. 132, pp. 1119–1128, 2018.
[15] S. Rajaraman and S. Antani, "Visualizing Convolutional Neural Network Decisions in Medical Image Classification," *Applied Sciences*, vol. 8, no. 9, pp. 1–17, 2018.
[16] A. Sandage, "The Classification of Galaxies: Early History and Ongoing Developments," *Annual Review of Astronomy and Astrophysics*, vol. 43, pp. 581–624, 2005.
[17] E. N. Lawrence et al., "Applications of Artificial Intelligence in Astronomical Image Processing," *Astronomy and Computing*, vol. 40, pp. 100–110, 2022.

APPENDIX

I. Project Repository
https://github.com/Jaishree-baskaran/Cosmoscan
II. Deployed Application
https://cosmoscan-galaxy-classifier.streamlit.app/