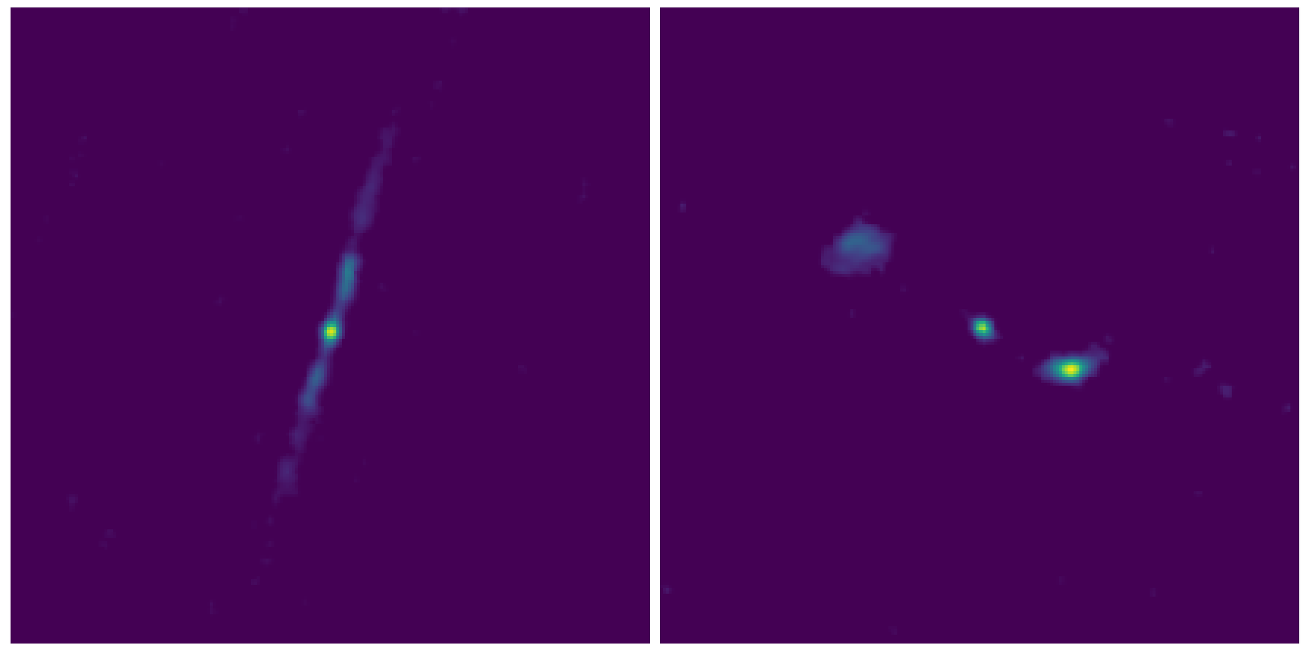
Morphological Classification of Radio Galaxies using Semi-Supervised Group Equivariant CNNs

Mir Sazzat Hossain   Sugandha Roy   K. M. B. Asad   Arshad Momen   Amin Ahsan Ali   Md Ashraful Amin   A. K. M. Mahbubur Rahman

[1]Center for Computational and Data Sciences (CCDS), Indepeddent University, Bangladesh

Radio Galaxies: Types and Morphology

Radio galaxies are specialized Active Galactic Nuclei (AGN), predominantly emitting radio waves. These waves result from charged particles accelerated by their central supermassive black holes. Notably, Sygnus A, discovered in 1937, remains the brightest detected source. Radio galaxies portray vast morphological diversities e.g. FRI, FRII, Head-tail, Ringlike, X-Shaped etc.



(a) Fanaroff-Riley Type-I (FRI)   (b) Fanaroff-Riley Type-II (FRII)

Figure 1. Examples of Fanaroff-Riley galaxies. (a) FRI galaxy, characterized by peak radio emission near the core and darker edges of the lobes. (b) FRII galaxy, characterized by peak radio emission at the edge of the lobes far from the core.

Importance and Challenges in Radio Galaxy Classification

Why Classification Matters:

- Provides insight into the formation and evolution of radio galaxies.
- FR classification reveals the AGN's power and radio emission distribution.
- Sheds light on the role of AGN in galactic evolution and interactions within its cosmic surroundings.
- Facilitates comparisons, trend identification, and hypothesis formulation about object nature.

The Daunting Challenge: With advancements in telescopes like the Square Kilometre Array (SKA), a flood of data is made available. This immense volume, combined with varied galaxy orientations, makes manual classification exceedingly challenging.

Our Innovative Contribution: In response to these challenges, our approach utilizes artificial intelligence, particularly the semi-supervised Group Equivariant Convolutional Neural Network (G-CNN). This ensures proficient classification, especially in scenarios where labeled data is scarce.

Related Deep-Learning Methods Used

Group Equivariant CNN (G-CNN): A specialized network maintaining equivariance to input symmetries, leveraging group convolutions to capture galaxy orientations effectively [4].

$$(f * \phi)(g) = \sum_{h \in G} f(h)\phi(h^{-1}g)$$

Contrastive Learning (SimCLR): Self-supervised, it learns by maximizing agreement between two views of the same data. It uses contrastive loss to ensure similarity between features of the same data while contrasting against different samples [1].

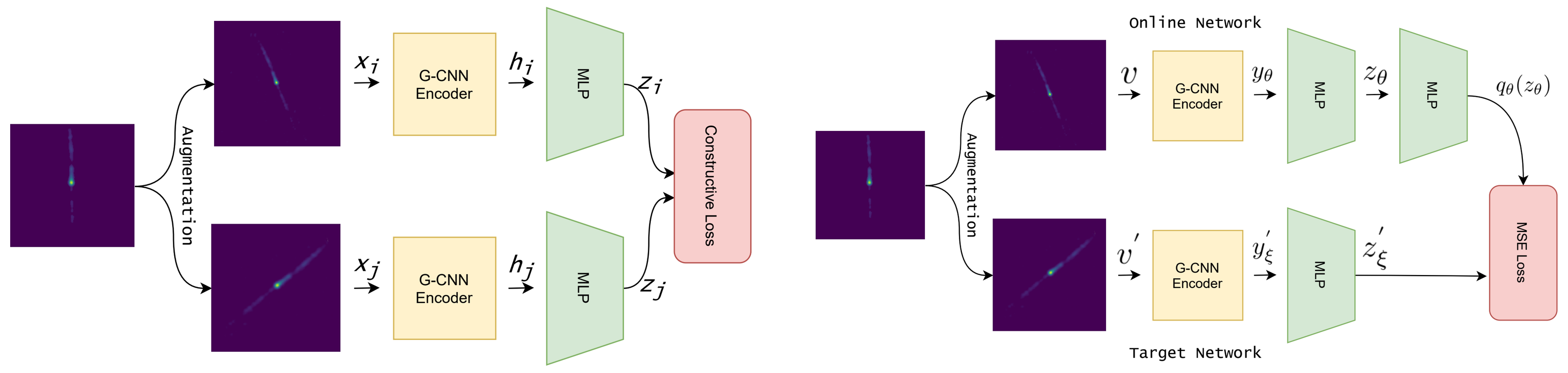
$$\mathcal{L}_{\text{SimCLR}} = -\log \frac{\exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^{2n} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k)/\tau)}$$

Bootstrap Your Own Latent (BYOL): A self-supervised method adopting "global contrastive learning". BYOL [2] trains an online and a target network, aiming for shared representations without needing explicit data augmentation or negative samples.

$$\mathcal{L}_{\theta,\xi} \triangleq \left\|\overline{q}_\theta(z_\theta) - \overline{z}'_\xi\right\|_2^2 = 2 - 2 \cdot \frac{\langle q_\theta(z_\theta), z'_\xi \rangle}{\|q_\theta(z_\theta)\|_2 \cdot \left\|z'_\xi\right\|_2}$$

Our Semi-Supervised Approach

Task-agnostic self-supervised learning:



(a) SimCLR   (b) BYOL

Figure 2. Illustration of (a) SimCLR and (b) BYOL for self-supervised learning. Replaced ResNet-50 with E(2)-Equivariant Steerable G-CNN.

Task-specific fine-tuning:

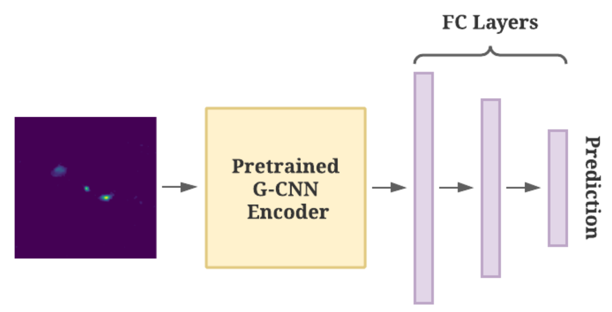


Figure 3. Illustration of mdoel used for fine-tuning and FR classification.

Datasets Used for Galaxy Classification

Our study primarily employs images from the Karl G. Jansky VLA's FIRST survey:

- Dataset-U: Sourced from the Radio Galaxy Zoo (RGZ), comprises 9,700 images (each 150×150 pixels) for self-supervised learning [5].
- Dataset-F: A subset of the MiraBest Batched Dataset [3], used for fine-tuning and FRI, FRII galaxy classifications.

Table 1. Numerical Identifiers of MiraBest Batched Dataset.

Digit 1	Digit 2	Digit 3
1 - FRI	0 - Confident	0 - Standard
2 - FRII	1 - Uncertain	1 - Double-double
3 - Hybrid		2 - Wide-angle Tail
		3 - Diffuse
		4 - Head-tail

Table 2. Source counts in Dataset-F.

Morphology	Train	Test	Total
FRI	348	49	397
FRII	381	55	436
Total	729	104	833

Results - Convergence Analysis

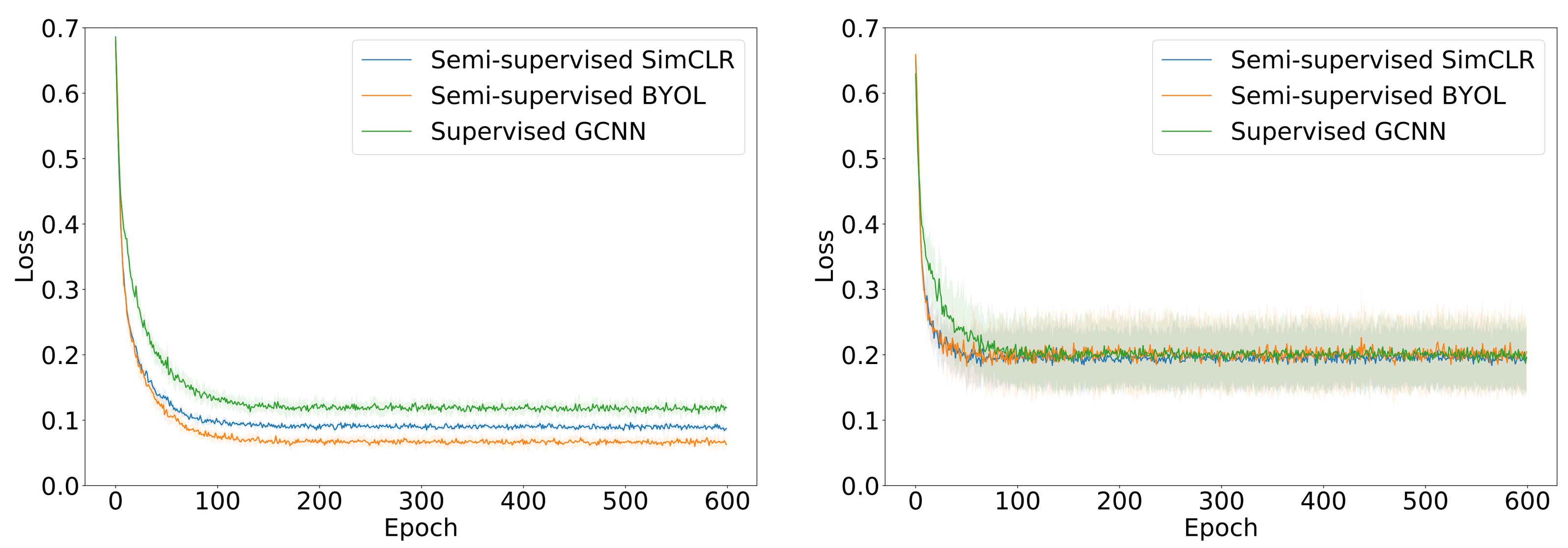


Figure 4. Convergence plots comparing fine-tuned SimCLR and BYOL encoders to a supervised G-CNN on Dataset-F. Displaying mean and standard deviation of (a) training loss and (b) validation loss over 5-fold cross-validation.

Results - Cluster Quality Analysis

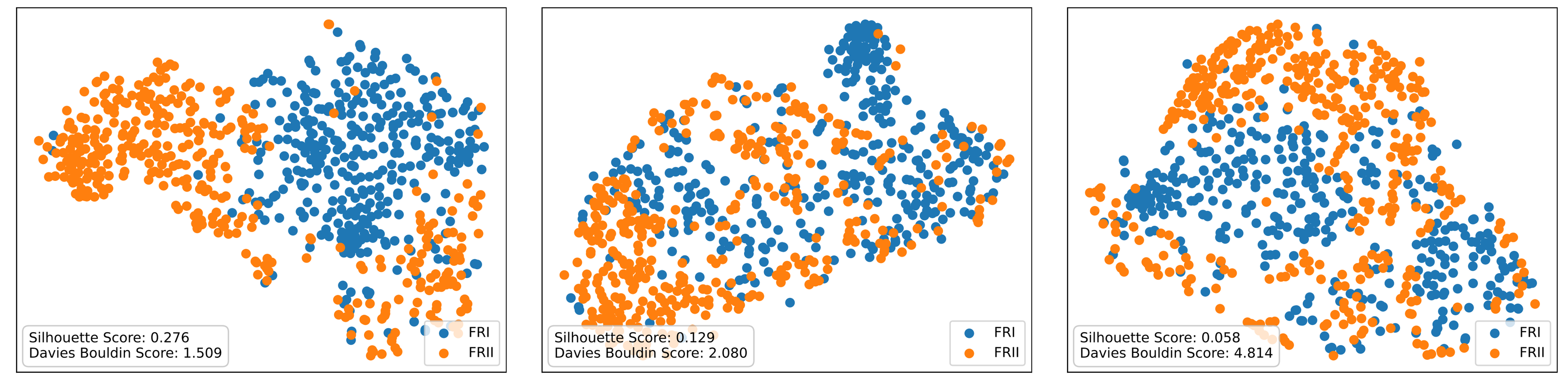


Figure 5. t-SNE visualizations of representations from the fine-tuned encoders of (a) BYOL, (b) SimCLR, and (c) Supervised G-CNN. Blue points denote FRI class, and orange for FRII. Improved clustering in our models, indicated by Silhouette and Davies Bouldin scores.

Results - Performance Comparison

Both models, BYOL and SimCLR, consistently outperformed in all classification metrics. The best metrics are highlighted in bold, and the second-best is underlined.

		FRI			FRII		
	Accuracy[%]	Precision	Recall	f1-score	Precision	Recall	f1-score
Semi-supervised SimCLR	$\underline{95.77 \pm 0.90}$	$\mathbf{0.98 \pm 0.061}$	0.93 ± 0.018	$\underline{0.95 \pm 0.011}$	0.94 ± 0.013	$\mathbf{0.98 \pm 0.014}$	$\underline{0.96 \pm 0.009}$
Semi-supervised BYOL	$\mathbf{97.12 \pm 0.40}$	$\underline{0.97 \pm 0.008}$	$\underline{0.96 \pm 0.009}$	$\mathbf{0.97 \pm 0.005}$	$\underline{0.96 \pm 0.007}$	$\underline{0.98 \pm 0.008}$	$\mathbf{0.97 \pm 0.004}$
Supervised G-CNN	94.80 ± 0.90	0.93 ± 0.012	$\mathbf{0.96 \pm 0.010}$	0.94 ± 0.009	$\mathbf{0.96 \pm 0.009}$	0.94 ± 0.012	0.95 ± 0.009

Table 3. Performance metrics comparing Semi-supervised models against Supervised methods. Our Semi-supervised approach demonstrates marked improvement.

Results - ROC Analysis and Statistical Significance

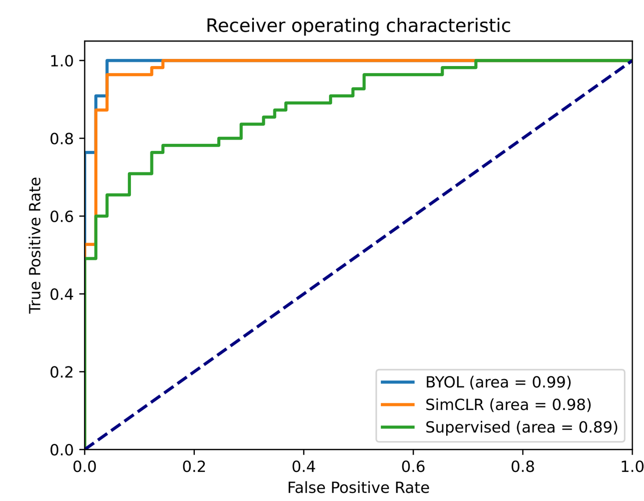


Figure 6. ROC curves for the fine-tuned encoders of BYOL, SimCLR, and the Supervised model, illustrating radio galaxy classification performance. Corresponding AUC scores are also provided.

Paired t-tests comparing our models to the supervised G-CNN showed significant improvements with SimCLR (p-value ≈ 0.08) and BYOL (p-value ≈ 0.0038), emphasizing the superiority of our semi-supervised approach.

References

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20, pages 1597–1607. JMLR.org, 2020.

[2] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

[3] Fiona Alice May Porter. Mirabest batched dataset, November 2020.

[4] Anna M. M. Scaife and Fiona Porter. Fanaroff-Riley classification of radio galaxies using group-equivariant convolutional neural networks. *Monthly Notices of the Royal Astronomical Society*, 503(2):2369–2379, May 2021.

[5] Chen Wu et al. Radio Galaxy Zoo: CLARAN - a deep learning classifier for radio morphologies. *Monthly Notices of the Royal Astronomical Society*, 482(1):1211–1230, January 2019.