RGC-BENT: A Novel Dataset for Bent Radio Galaxy Classification

Mir Sazzat Hossain[1], Khan Muhammad Bin Asad[2], Payaswini Saikia[3], Adrita Khan[1], Md Akil Raihan Iftee[1], Rakibul Hasan Rajib[1], Arshad Momen[1], M. Ashraful Amin[1], Amin Ahsan Ali[1], A.K.M. Mahbubur Rahman[1]

1. Center for Computational & Data Sciences, Independent University, Bangladesh
2. Center for Astronomy, Space Science and Astrophysics, Independent University, Bangladesh
3. Center for Astrophysics and Space Science, New York University Abu Dhabi

2025 IEEE International Conference on Image Processing

Presenter: Mir Sazzat Hossain

Introduction to Bent Radio AGN

Active Galactic Nuclei (AGN): Luminous objects powered by accretion onto supermassive black holes, crucial for understanding galaxy evolution.

Bent Radio AGN: A subset of radio-loud AGN characterized by distinctly curved jet structures due to interactions with the intergalactic medium (ICM).

- Commonly observed in galaxy clusters.
- Bending often linked to ram pressure from relative motion between host galaxy and ICM.

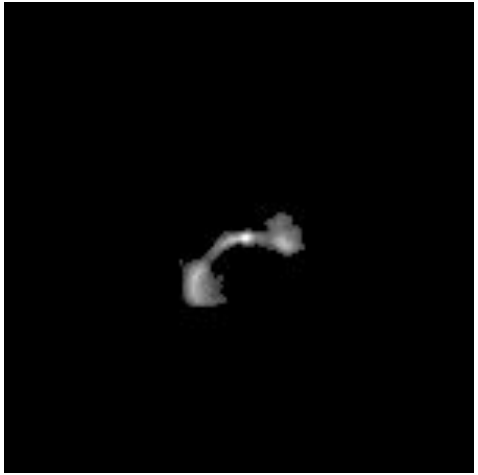


Fig. 1 Sample Bent Radio AGN

Introduction to Bent Radio AGN

Two Subtypes:

- **Wide-Angle Tailed (WAT) Galaxies:** Two-sided jets forming a wide 'C' shape with an opening angle exceeding 90°.
- **Narrow-Angle Tailed (NAT) Galaxies:** Closely aligned jets forming a narrow 'V' shape with an opening angle under 90°.

Significance: Studying bent radio AGN provides insights into galaxy cluster dynamics, ICM behavior, galaxy-environment interactions, and AGN jet dynamics.

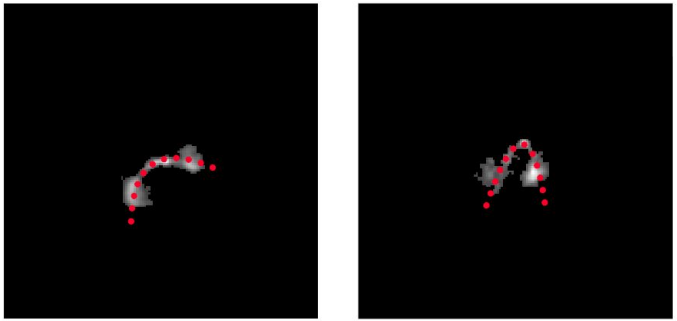


Fig. 2: Examples of WAT and NAT radio galaxies

Limitations of Prior Works

- **Scarcity of Data**: Lack of specialized datasets and benchmarks for bent radio AGN.
- **Existing Datasets**:
  - ***Galaxy Zoo [1]****:* Focuses on general galaxy classification.
  - ***Radio Galaxy Zoo [2]****:* Lacks specific WAT and NAT labeling.
  - ***MiraBest [3]****:* Provides labeled Fanaroff-Riley Type I and II sources, but not WATs and NATs.
- **Previous Catalogs**: *Proctor et al. [4]* contains unreliable automated annotations.
- **Our Data Source**: The *Sasmal et al. [5]* catalog requires extensive preprocessing and validation to be machine learning-ready.

Key Contributions

- Introduction of a novel machine learning dataset of bent radio AGN images (from Very Large Array - VLA).
- Detailed data acquisition and processing steps, including expert validation.
- Evaluation of state-of-the-art deep learning models (CNNs and transformers) on this dataset, establishing a baseline.
- Release of all source code for key data preprocessing tasks (background estimation, source identification, mask generation, background removal).

Dataset Creation: Data Acquisition

- **Raw Data**: Sourced from the Faint Images of the Radio Sky at Twenty-cm (FIRST) survey using the VLA telescope.
- **Source Coordinates**: Obtained from the *Sasmal et al.* catalog.
- **Data Format**: Images are downloaded from the NASA SkyView virtual observatory in Flexible Image Transport System (FITS) format.
- **Key Specs**:
  - **Frequency**: 1.4 GHz
  - **Resolution**: 5-arcsecond
  - **Sensitivity**: 0.15 mJy

Fig. 3: VLA Telescope

Dataset Creation: Data Processing

1. **Background Estimation**

   Goal: To figure out the average "noise level" in each image.
   We calculate the **local background level** using this formula:

$$B(x, y) = \mu_{\text{local}} + k \cdot \sigma_{\text{local}}$$

   - $B(x, y)$: background intensity at pixel (x, y)
   - μ_{local}: average brightness in a small region
   - σ_{local}: variation in brightness (standard deviation)
   - k: a scaling factor to account for noise (e.g., 3 or 5)

   This helps us distinguish between background noise and actual galaxy signals.

Dataset Creation: Data Processing

2. Source Identification

Goal: To detect the bright regions (potential galaxies) in the image.

We use thresholds to find regions that stand out:

- **Island Threshold:**
  $$I(x, y) > B(x, y) + T_{\text{isl}} \cdot \sigma_{\text{local}}$$
- **Peak Threshold:**
  $$I(x, y) > B(x, y) + T_{\text{pix}} \cdot \sigma_{\text{local}}$$
- $I(x, y)$: pixel intensity
- T_{isl}: minimum level for a region (set to 3)
- T_{pix}: minimum level for a peak (set to 5)

This allows us to identify and isolate potential bent AGN sources.

Dataset Creation: Data Processing

3. Modeling the Galaxy Shape

Goal: To mathematically describe each galaxy's appearance.

We fit a 2D Gaussian model to each detected source:

$$I(x, y) = I_0 \cdot \exp\left(-\left(\frac{(x' - x_0)^2}{2\sigma_{\text{maj}}^2} + \frac{(y' - y_0)^2}{2\sigma_{\text{min}}^2}\right)\right)$$

With the corresponding variable definitions:

- I_0: peak brightness
- (x_0, y_0): center coordinates
- $\sigma_{\text{maj}}, \sigma_{\text{min}}$: major and minor axis widths
- θ: angle of rotation (used to calculate x', y')

This helps us extract meaningful features like size, orientation, and intensity.

Dataset Creation: Data Processing

4. Mask Generation

Goal: To create a binary mask that delineates the precise spatial extent of each source. A binary mask, M(x,y), is generated by thresholding the fitted Gaussian model using the equation:

$$M(x,y) = \begin{cases} 1, & \text{if } I(x,y) > B(x,y) + T_{\text{pix}} \cdot \sigma_{\text{local}}, \\ 0, & \text{otherwise.} \end{cases}$$

Dilation: For extended or diffuse sources, a dilation operation is applied to expand the mask boundary and ensure complete coverage.

Dataset Creation: Data Processing

5. Image Conversion (FITS to PNG)

Goal: To prepare the images for visualization and deep learning models, we scale them from radio signal values to pixel brightness (0–255):

$$\text{img}_{\text{PNG}} = \left(\frac{\text{img} - \min(\text{img})}{\max(\text{img}) - \min(\text{img})} \right) \cdot 255$$

Dataset Creation: Data Processing

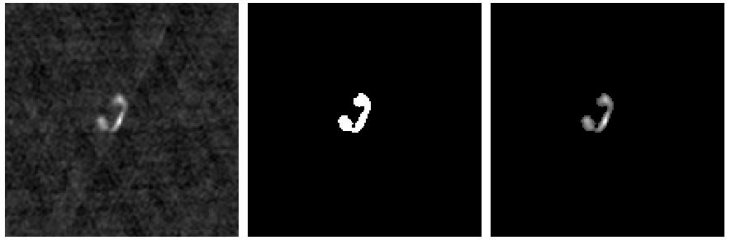
6. Background Removal

We clean the images by applying a binary mask:

$$\text{img}_{\text{masked}} = \text{img} \times M$$

With the corresponding definitions:

- M: the mask (1 = keep pixel, 0 = remove pixel)
- $\text{img}_{\text{masked}}$: cleaned image with only the galaxy



(a) Original Image (b) Generated Mask (c) Masked Image

Fig. 4: Background Removal Process

Dataset Creation: Final Dataset Curation

Expert Validation:

- Two expert astronomers manually reviewed each processed image.
- Verified presence of bent AGN, mask validity, background removal quality, and overall image quality.
- From a total of 703 sources, 64 sources got rejected.
- Resulting in a final dataset of 639 bent radio AGNs.

Dataset Creation: Final Dataset Curation

Dataset Splitting:

- Split into 10 batches (9 for training, 1 for testing).
- Each batch contains 64 images, except the last (63 images).
- Stratified based on source type (NAT or WAT) to ensure even distribution.
- Stored as separate pickle files.
- **Training Set:** First 9 batches (576 images)
- **Test Set:** Last batch (63 images)
- **Total:** 639 images (254 NAT, 385 WAT)

Batch No.	Source Count		Total Sources	Cumulative Total (Train/Test)
	WAT	NAT		
Training Set				
0	39	25	64	
1	39	25	64	
2	39	25	64	
3	39	25	64	
4	39	25	64	576
5	38	26	64	
6	38	26	64	
7	38	26	64	
8	38	26	64	
Test Set				
9	38	25	63	63
Total	385	254	639	639

Tab. 1 Distribution of NAT and WAT sources in RGC-Bent

Benchmarking Experiments:

Experimental Setup:

- **Baseline Models:** VGG-16, ResNet-50 (CNNs), Vision Transformer (ViT-B-16), SWIN Transformer (SWIN-B) (Transformer-based), ConvNeXT (CNN-based).
- All models pre-trained on ImageNet and fine-tuned on our dataset.
- Input images: grayscale 150×150 pixels, resized to 224×224 and converted to RGB (duplicated grayscale channel).
- Optimizers: Adam (CNNs), AdamW (Transformers).
- Batch size: 32; Epochs: 20; Early stopping.

Evaluation Metrics: Accuracy, Precision, Recall, F1-score (class-wise for NAT and WAT).

Quantitative Results

Model	Accuracy [%]	WAT			NAT		
		Precision	Recall	F1-score	Precision	Recall	F1-score
VGG-16	74.60	0.73	0.92	0.81	0.80	0.48	0.60
ResNet-50	77.77	0.79	0.87	0.82	0.76	0.64	0.70
ViT-B-16	76.19	0.76	0.89	0.82	0.77	0.73	0.74
SWIN-B	80.95	0.80	0.92	0.85	0.84	0.64	0.73
ConvNeXT	84.12	0.85	0.89	0.87	0.83	0.76	0.79

Tab. 2 Performance comparison of different deep learning models on RGC-Bent

Key Finding: ConvNeXT achieved the highest F1-scores for both WAT (0.87) and NAT (0.79) sources, demonstrating its robustness.

Qualitative Analysis

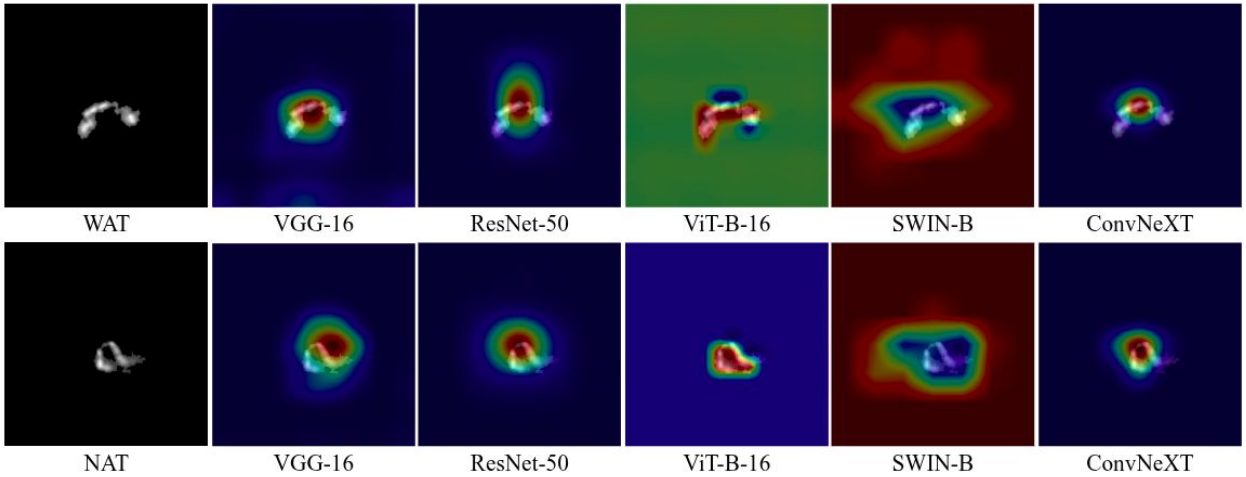


Fig. 5 CAMs generated by different models for a sample NAT and WAT source.

Key Finding: CNNs focus on the core regions of the source. Transformer-based models capture the elongated structure more effectively. This explains why transformers show higher recall for WAT sources.

Conclusion

- Introduced RGC-BENT, a novel, expertly curated dataset of 639 bent radio AGN images (WAT and NAT).
- Provided detailed data acquisition, processing, and validation methodologies.
- Benchmarked various deep learning models, showing the effectiveness of advanced machine learning for bent radio AGN classification.
- ConvNeXT emerged as the top performer, achieving the highest F1-scores for both WAT and NAT sources.
- Dataset and code are open-source, facilitating further research.

Thank You

References

[1] Chris J. Lintott, Kevin Schawinski, An˘ze Slosar, Kate Land, Steven Bamford, Daniel Thomas, M. Jordan Raddick, Robert C. Nichol, Alex Szalay, Dan Andreescu, Phil Murray, and Jan Vandenberg, "Galaxy zoo: mor- phologies derived from visual inspection of galaxies from the sloan digital sky survey," Monthly Notices of the Royal Astronomical Society, vol. 389, no. 3, pp. 1179–1189, 09 2008.*

[2] O. Ivy Wong, A. F. Garon, M. J. Alger, L. Rudnick, S. S. Shabala, K. W. Willett, J. K. Banfield, H. Andernach, R. P. Norris, J. Swan, M. J. Hardcastle, C. J. Lintott, S. V. White, N. Seymour, A. D. Kapi´nska, H. Tang, B. D. Simmons, and K. Schawinski, "Radio Galaxy Zoo data release 1: 100185 radio source classifications from the FIRST and ATLAS surveys," Monthly Notices of the Royal Astronomical Society, vol. 536, no. 4, pp. 3488–3506, Feb. 2025.

[3] Fiona A M Porter and Anna M M Scaife, "Mirabest: a data set of morphologically classified radio galaxies for machine learning," RAS Techniques and Instruments, vol. 2, no. 1, pp. 293–306, 06 2023.

[4] D Proctor, "Morphological annotations for groups in the first database," The Astrophysical Journal Supplement Series, vol. 194, no. 2, pp. 31, June 2011.

[5] apan K. Sasmal, Soumen Bera, Sabyasachi Pal, and Soumen Mondal, "A New Catalog of Head-Tail Radio Galaxies from the VLA FIRST Survey," The Astrophysical Journal Supplement Series, vol. 259, no. 2, pp. 31, Apr. 2022.