Data in Brief 47 (2023) 108974

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

FIRST radio galaxy data set containing curated labels of classes FRI, FRII, compact and bent

Florian Griese[a,b,c,*], Janis Kummer[a,d], Patrick L.S. Connor[a,e], Marcus Brüggen[a,d], Lennart Rustige[a,f]

[a] Center for Data and Computing in Natural Sciences (CDCS), Notkestrasse 9-11, D-22607 Hamburg, Germany
[b] Institute for Biomedical Imaging, Hamburg University of Technology, D-21073 Hamburg, Germany
[c] Section for Biomedical Imaging, University Medical Center Hamburg-Eppendorf, D-20246 Hamburg, Germany
[d] Universität Hamburg, Hamburger Sternwarte, Gojenbergsweg 112, D-21029 Hamburg, Germany
[e] Institut für Experimentalphysik, University of Hamburg, Luruper Chaussee 149, D-22761 Germany
[f] Deutsches Elektronen-Synchrotron DESY, Notkestrasse 85, Hamburg D-22607, Germany

ARTICLE INFO

Article history:
Received 23 November 2022
Revised 2 February 2023
Accepted 7 February 2023
Available online 11 February 2023

Dataset link: RadioGalaxyDataset (Original data)

Keywords:
Radio Galaxy
FIRST survey
FRI
FRII
Bent
Compact
Fanaroff-Riley

ABSTRACT

Automated classification of astronomical sources is often challenging due to the scarcity of labelled training data. We present a data set with a total number of 2158 data items that contains radio galaxy images with their corresponding morphological labels taken from various catalogues [1,2]. The data set is curated by removing duplicates, ambiguous morphological labels and by different meta data formats. The image data was acquired by the VLA FIRST (Faint Images of the Radio Sky at Twenty-Centimeters) survey [3]. The morphological labels are collected and the catalogue specific classification definition is converted into a 4-class classification scheme: FRI, FRII, Compact and Bent sources. FRI and FRII correspond to the two classes of the widely used Faranoff-Riley classification [4]. We consider two more classes: compact sources and bent-tail galaxies. For duplicates with different morphological labels, the galaxy is regarded as ambiguously labeled and both coordinates are removed. For the remaining list of coordinates, the radio galaxy images are collected from the virtual observatory skyview (https://skyview.gsfc.nasa.gov/current/cgi/query.pl). The gray value images are provided in the size of 300 × 300 pixel and all pixels with a

* Corresponding author.
E-mail address: florian.griese@tuhh.de (F. Griese).
Social media: cdcs_hamburg (F. Griese)

https://doi.org/10.1016/j.dib.2023.108974

value below three times the local RMS of the noise are set to this threshold value. The data set is useful for the development of robust machine learning models that automate the classification of radio galaxy images.

Specifications Table

Subject	Astronomy and Astrophysics
Specific subject area	Radio astronomy: Morphological Classification of radio galaxy images
Type of data	Image with class label
How data were acquired	Images are collected by coordinate from virtual observatory skyview (https://skyview.gsfc.nasa.gov/current/cgi/query.pl) and labels are collected by coordinate from catalogues
Data format	Raw (png and csv), Filtered (hdf5)
Description of data collection	The radio galaxy images are collected by sending a html request to the virtual observatory skyview (https://skyview.gsfc.nasa.gov/current/cgi/query.pl) with the following parameters: coordinates, sigma = 3 and image patch size = 300. The website returns gray images with uInt8 data type normalized between 0 and 255. Corresponding to coordinates, the morphological labels are published by the catalogues in supplementary files in form of text files (.txt) or PDFs. Only images from the four classes FRI, FRII, Compact and Bent are considered for the data set.
Data source location	Primary data sources: Radio galaxy images from VLA FIRST (1.4GHz) survey [3]: • Virtual observatory skyview (https://skyview.gsfc.nasa.gov/current/cgi/query.pl) VLA FIRST (1.4 GHz) survey **Provenance**: The FIRST project team: R.J. Becker, D.H. Helfand, R.L. White M.D. Gregg. S.A. Laurent-Muehleisen. **Copyright:** 1994, University of California. Permission is granted for publication and reproduction of this material for scholarly, educational, and private non-commercial use. Inquiries for potential commercial uses should be addressed to: Robert Becker, Physics Dept, University of California, Davis, CA 95616 Institution: NASA City/Town/Region: CA Country: U.S. Latitude and longitude (and GPS coordinates, if possible) for collected samples/data: Label information: • Gendre [5,6] Catalogue, Table: mnras0404-1719-SD1.pdf, data tables CoNFIG-1 to CoNFIG-4 • Capetti [7,8], – Catalogue,Table – Catalogue, Table • Baldi [9], Catalogue, Table • Mira Best [10], Catalogue, Table • Proctor [11], Catalogue, Table, data from Table 1 from reference with label “WAT” and “NAT”
Data accessibility	Repository name: RadioGalaxyDataset Data identification number: 10.5281/zenodo.7351724 [2] Direct URL to data: https://doi.org/10.5281/zenodo.7351724.
Related research article	J. Kummer, L. Rustige, F. Griese, K. Borras, M. Brüggen, P. L. S. Connor, F. Gaede, G. Kasieczka, P. Schleper, Radio galaxy classification with wgan-supported augmentation, in: INFORMATIK 2022, volume P-326 of Lecture Notes in Informatics (LNI) - Proceedings, Gesellschaft für Informatik, Bonn, 2022, pp. 469-478. doi:10.18420/inf2022_38. [1]

Value of the Data

- The data set is useful as it provides an easy to access, curated and combined data set based on various catalogues. This data set can be used to develop supervised deep learning models to classify radio galaxies in the categories FRI, FRII, Compact and Bent.
- Computer scientists in the field of Astronomy and Astrophysics who are developing supervised, self-supervised or unsupervised deep learning models for automatic classification, object detection or data generation. Further, the data set can be used to validated and evaluated unsupervised models.
- In combination with a labeled LOFAR data set, the data set can be used to develop a model that generalizes to data from another telescope operating in a different wavelength range.
- The easy-accessible and curated data set is suitable for educational purposes in applying machine learning methods to astronomical data.

1. Objective

The data set is created to train supervised deep learning models on radio galaxy data. However, the available data set [12] has a limited number of 1256 data entries. The extraction of data from various catalogues turned out to be challenging because meta data is not consistent. In some catalogues, identical radio galaxy sources have different coordinates due to different rounding schemes. Further, identical radio galaxy sources can have different classification labels between different catalogues. In this data set, data items can easily be filtered by class label, catalogue or coordinate range. Researchers should have the ability to build on this data set and do not have to repeat this work.

2. Data Description

We combined different catalogues which characterise radio galaxy sources from the FIRST survey [3] to create a data set of radio galaxy images with morphological labels. The labeling is typically done by experts by considering radio images and the corresponding optical counterparts. In this work, we group radio sources in 4 classes (FRI, FRII, Compact and Bent) as done in [13,14]. FRI and FRII are defined by Fanaroff-Riley in [4]. The Compact class consists of unresolved point sources. The Bent class consists of sources for which the angle between the jets differs significantly from 180 degrees. It contains two subtypes: narrow-angle tail (NAT) and wide-angle tail (WAT) galaxies, depending on the angle between the jets. In Fig. 1 adapted from [1], a few examples of each class are shown. The created data set has a total size of 5.1 MB.

2.1. firstgalaxydata/galaxy_data.zip

The data set can be found in the galaxy_data folder by unzipping galaxy_data.zip. It contains the folder structure

- /[all,train,valid,test]/FRI
- /[all,train,valid,test]/FRII
- /[all,train,valid,test]/Compact
- /[all,train,valid,test]/Bent

with corresponding.png images. The most import information will also be part of the file name separated by underscores: ‘RA_DEC_Label_Catalogue.png‘ E.g. ‘14.084_-9.608_3_MiraBest.png‘

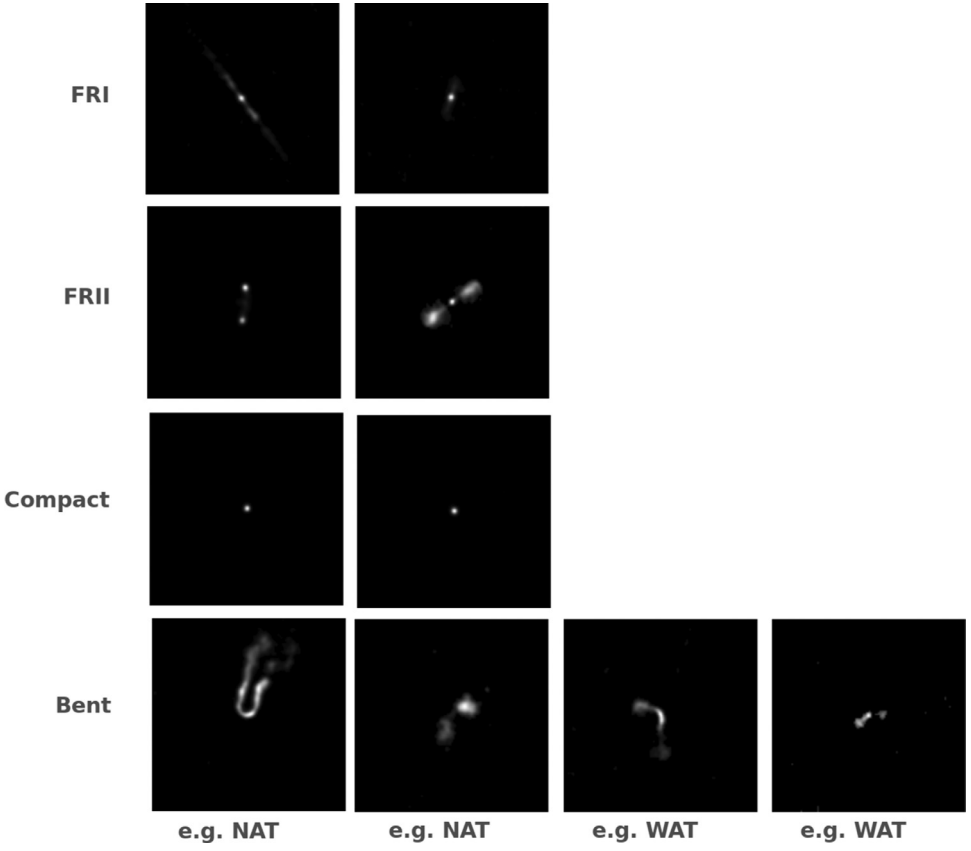

Fig. 1. Examples of the classes FRI, FRII, Compact and Bent. The figure is adapted from the original figure in [1].

Table 1
Number of sources per class in the data set. The table is adapted from the original table in [1].

	FRI	FRII	Compact	Bent	Total (%)
train	395	824	291	248	1758 (81,46%)
validation	50	50	50	50	200 (9,26%)
test	50	50	50	50	200 (9,26%)
total (%)	495 (22,93%)	924 (42,81%)	391 (18,11%)	348 (16,12%)	2158 (100%)

The number of radio galaxy sources per class and split are given in Table. 1.

2.2. *firstgalaxydata/galaxy_data_h5.zip*

The combined data set collected from the FIRST catalogues is summarized in the HDF5 file galaxy_data_h5.h5 with a group named "data_$(i)" for every data entry with $i = 1, \ldots, n$ with n as the total number of data entries. Each group has the following data sets

- "Img": two-dimensional uint8 array with (300,300), The data set "Img" has the following attributes

Table 2
Re-labeled galaxies from FRICat Capetti catalogue [7].

	RA in ◦	DEC in ◦	Original Label	Corrected Label
1	140.204	40.665	FRI	Bent
2	152.017	50.445	FRI	Bent
3	159.613	41.815	FRI	Bent
4	181.006	20.232	FRI	Bent
5	181.105	3.753	FRI	Bent
6	216.568	0.838	FRI	Bent
7	223.064	50.374	FRI	Bent
8	227.489	33.454	FRI	Bent
9	250.225	32.791	FRI	Bent

  - "RA": double, right ascension equatorial coordinate system (J2000)
  - "DEC" double, declination equatorial coordinate system (J2000)
  - "Source": string ["Gendre", "MiraBest", "Capetti2017a", "Capetti2017b", "Baldi2018", "Proctor_Tab1"]
  - "Filepath_literature": string, relative path to the *.png file in folder 'galaxy_data'
- "Label_literature": double, 0.0: "FRI", 1.0: "FRII", 2.0: "Compact", 3.0: "Bent"
- "Split_literature" : string, ["train", "valid", "test"]

2.3. *firstgalaxydata/firstgalaxydata.py*

This python class is able to load the galaxy_data_h5.h5 with several filtering options to provide a data.Dataset class for the pytorch framework.

2.4. *firstgalaxydata/Example_firstgalaxydata.py*

This file shows example code on how to use the firstgalaxydata.py class.

2.5. *requirements.txt*

The requirements.txt contains the necessary packages in order to use the firstgalaxydata.py class with python.

2.6. *meta/FRICat_Capetti_2017_relabeled.csv*

The FRICat_Capetti_2017_relabeled.csv contains the relabeled sources from FRICat Capetti catalogue as shown in Table. 2.

2.7. *meta/galaxy_data_removed.csv*

The galaxy_data_removed.csv file contains a list of sources that are not included in the data set. These sources are already added to the data set but have slightly different coordinates within the deviation of $\pm 0.015^{\circ}$ in right ascension (RA) and declination (DEC). Sources with different coordinates within the deviation are regarded as duplicates.

2.8. meta/galaxy_data_different_labels.csv

The galaxy_data_different_labels.csv contains a list of pairs of coordinates which are within the deviation of $\pm 0.015^{\circ}$ in right ascension (RA) and declination (DEC) but in this case the label information is different from the original catalogues. These source coordinates are entirely excluded from the data set.

2.9. img/Classification_Scheme.png

An image showning examples of the four classes.

3. Experimental Design, Materials and Methods

The morphological labels are collected and assigned to the corresponding class from the following catalogues. For mapping the labels to the radio galaxy images, the equatorial coordinates (J2000) with right ascension (RA) and declination (DEC) are used. For the FRI class, we used the catalogue of [5,6] by collecting images from data tables CoNFIG-1 to CoNFIG-4 with label "I". Additionaly, images from the catalogue of [10] were collected with label "0" and the images from the catalogue of [7] were selected. For the FRII class, we used the catalogue of [5,6] by collecting images from data tables CoNFIG-1 to CoNFIG-4 with label "II" and "IIc". The catalogue of [10] with images with label "5" and "6" provided further images of FRII along with the images of the catalogue of [8]. For the Compact class, we used the catalogue of [5,6] by collecting images from data tables CoNFIG-1 to CoNFIG-4 with label "C", "C*" and "S*". Further, the catalogue of [9] was used for Compact. For the Bent class, we used the catalogue of [5,6] by collecting images from data tables CoNFIG-1 to CoNFIG-4 with label "Iw". Further, we selected bent-type sources from [11] by collecting only from Table 1 from [11] with label "WAT" and "NAT". From catalogue [10] we collected images with label "1" and "2" for the Bent class.

We identified 300 source coordinates within a deviation of $\pm 0.015^{\circ}$ in right ascension (RA) and declination (DEC) from different catalogues. 227 of 300 source coordinates had the same label information from different catalogues. Here, the radio galaxy image is only added once to data set and the listed 227 source coordinates are regarded as duplicates. 73 of 300 source coordinates had different label information from different catalogues. Here, the radio galaxy images are ambiguously labeled and excluded from the data set entirely. From the FRICAT catalogue [7] 9 sources were manually re-labeled from FRI to Bent since these sources are NAT galaxies. The coordinates of the re-labeled galaxies are listed in Table. 2.

3.1. Preprocessing

We downloaded the images of the FIRST survey from the virtual observatory skyview (https://skyview.gsfc.nasa.gov) using the equatorial coordinates (J2000). The original images size is 300 x 300 pixel. We adopted the preprocessing and the choice of preprocessing parameters from [15] and [13]. At first, all NaN value are set to zero. Second, all pixel values below three times the local RMS noise are set to the value of this threshold with help of the following functions sigma_clipped_stats and preprocess_clip_normalize. The value of sigma equal to 3 ensures that

the background noise of the images is cleared. We have used python version 3.6.8, numpy version 1.19.5 and astropy version 4.1.

```
import numpy as np

def preprocess_clip_normalize(img, std, sigma=3.0):
    """ Clips the image with respect to sigma * standard deviation.
    :param img: numpy.ndarray (float)
        image
    :param std: float
        standard deviation
    :param sigma: float
        factor of how many std should be clipped
    :return: numpy.ndarray (float)
        normalized image between 0 and 1
    """
    # clip with sigma * standard deviation
    img_clip = np.clip(img, sigma * std, np.inf)
    # normalize to 0 and 1
    img_norm = (img_clip - np.min(img_clip)) / (np.max(img_clip) - np.min(img_clip))
    return img_norm


def convert_to_unit8(img):
    """
    Convert to range 0 to 255 and uint8 in order to make convertible to PIL object
    :param img: numpy.ndarray (float)
        img should be normalized between 0 and 1
    :return: numpy.ndarray (uint8)
        img is between 0 and 255
    """
    img = 255 * img
img = img.astype(np.uint8)
return img
```

Ethics Statements

The authors comply with redistribution polices of the primary data sources and have stated the license for the reuse of primary data.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships which have, or could be perceived to have, influenced the work reported in this article.

Data Availability

RadioGalaxyDataset (Original data) (Zenodo).

CRediT Author Statement

Florian Griese: Conceptualization, Methodology, Data curation, Software, Writing – original draft; **Janis Kummer:** Conceptualization, Data curation, Validation, Writing – review & editing; **Patrick L.S. Connor:** Writing – review & editing; **Marcus Brüggen:** Writing – review & editing, Supervision, Funding acquisition; **Lennart Rustige:** Data curation, Validation, Writing – review & editing.

Acknowledgments

This work was supported in part by UHH, DESY, TUHH and HamburgX grant LFF-HHX-03 to the Center for Data and Computing in Natural Sciences (CDCS) from the Hamburg Ministry of Science, Research, Equalities and Districts.

References

[1] J. Kummer, L. Rustige, F. Griese, K. Borras, M. Brüggen, P.L.S. Connor, F. Gaede, G. Kasieczka, P. Schleper, Radio galaxy classification with wGAN-supported augmentation, in: Informatik 2022, in: Lecture Notes in Informatics (LNI) - Proceedings, vol. P-326, Gesellschaft für Informatik, Bonn, 2022, pp. 469–478, doi:10.18420/inf2022_38.
[2] Griese, F. (2022). doi:10.5281/zenodo.7351724.
[3] R.H. Becker, R.L. White, D.J. Helfand, The FIRST Survey: Faint Images of the Radio Sky at Twenty Centimeters, Astrophys. J. 450 (1995) 559, doi:10.1086/176166. https://skyview.gsfc.nasa.gov/current/cgi/query.pl
[4] B.L. Fanaroff, J.M. Riley, The morphology of extragalactic radio sources of high and low luminosity, MNRAS 167 (1974) 31P–36P, doi:10.1093/mnras/167.1.31P.
[5] M.A. Gendre, J.V. Wall, The combined NVSS-FIRST galaxies (coNFIG) sample - i. sample definition, classification and evolution, Month. Notice. R. Astron. Soc. (2008), doi:10.1111/j.1365-2966.2008.13792.x.
[6] M.A. Gendre, P.N. Best, J.V. Wall, The combined NVSS-FIRST galaxies (coNFIG) sample - II. comparison of space densities in the fanaroff-riley dichotomy, Month. Notice. R. Astrono. Soc. (2010), doi:10.1111/j.1365-2966.2010.16413.x.
[7] A. Capetti, F. Massaro, R.D. Baldi, Fricat: A first catalog of fr i radio galaxies, Astron. Astrophys. 598 (2017) A49, doi:10.1051/0004-6361/201629287.
[8] A. Capetti, F. Massaro, R.D. Baldi, Friicat: A first catalog of fr ii radio galaxies, Astron. Astrophys. 601 (2017) A81, doi:10.1051/0004-6361/201630247.
[9] R.D. Baldi, A. Capetti, F. Massaro, Fr0cat: a first catalog of fr 0 radio galaxies, Astron. Astrophys. 609 (2017) A1, doi:10.1051/0004-6361/201731333.
[10] H. Miraghaei, P.N. Best, The nuclear properties and extended morphologies of powerful radio galaxies: the roles of host galaxy and environment, Month. Notice. R. Astron. Soc. (2017) stx007, doi:10.1093/mnras/stx007.
[11] D.D. Proctor, Morphological annotations for groups in the first database, Astrophys. J. Suppl. Ser. 194 (2) (2011) 31, doi:10.1088/0067-0049/194/2/31.
[12] Porter, F. (2020). doi:10.5281/zenodo.4288837.
[13] A. Samudre, L.T. George, M. Bansal, Y. Wadadekar, Data-efficient classification of radio galaxies, MNRAS 509 (2) (2021) 2269–2280, doi:10.1093/mnras/stab3144.
[14] W. Alhassan, A.R. Taylor, M. Vaccari, The FIRST classifier: compact and extended radio galaxy classification using deep convolutional neural networks, Month. Notice. R. Astronomic. Soc. 480 (2) (2018) 2085–2093, doi:10.1093/mnras/sty2038.
[15] A.K. Aniyan, K. Thorat, Classifying radio galaxies with the convolutional neural network, Astrophys. J. Suppl. Ser. 230 (2) (2017) 20, doi:10.3847/1538-4365/aa7333.