*Publications of the Astronomical Society of Australia* (2024), **41**, e085, 19 pages
doi:10.1017/pasa.2024.84

CAMBRIDGE UNIVERSITY PRESS

**Research Article**

# Self-supervised contrastive learning of radio data for source detection, classification and peculiar object discovery

S. Riggi[1], T. Cecconello[1,2], S. Palazzo[2], A.M. Hopkins[3], N. Gupta[4], C. Bordiu[1], A. Ingallinera[1], C. Buemi[1], F. Bufano[1], F. Cavallaro[1], M.D. Filipović[5], P. Leto[1], S. Loru[1], A.C. Ruggeri[1], C. Trigilio[1], G. Umana[1] and F. Vitello[1]

[1]INAF – Osservatorio Astrofisico di Catania, Via Santa Sofia 78, 95123 Catania, Italy, [2]Department of Electrical, Electronic and Computer Engineering, University of Catania, Viale Andrea Doria 6, 95125 Catania, Italy, [3]School of Mathematical and Physical Sciences, 12 Wally's Walk, Macquarie University, NSW 2109, Australia, [4]CSIRO Space & Astronomy, PO Box 1130, Bentley WA 6102, Australia and [5]Western Sydney University, Locked Bag 1797, Penrith South DC, NSW 2751, Australia

## Abstract

New advancements in radio data post-processing are underway within the Square Kilometre Array (SKA) precursor community, aiming to facilitate the extraction of scientific results from survey images through a semi-automated approach. Several of these developments leverage deep learning methodologies for diverse tasks, including source detection, object or morphology classification, and anomaly detection. Despite substantial progress, the full potential of these methods often remains untapped due to challenges associated with training large supervised models, particularly in the presence of small and class-unbalanced labelled datasets.

Self-supervised learning has recently established itself as a powerful methodology to deal with some of the aforementioned challenges, by directly learning a lower-dimensional representation from large samples of unlabelled data. The resulting model and data representation can then be used for data inspection and various downstream tasks if a small subset of labelled data is available.

In this work, we explored contrastive learning methods to learn suitable radio data representations by training the SimCLR model on large collections of unlabelled radio images taken from the ASKAP EMU and SARAO MeerKAT GPS surveys. The resulting models were fine-tuned over smaller labelled datasets, including annotated images from various radio surveys, and evaluated on radio source detection and classification tasks. Additionally, we employed the trained self-supervised models to extract features from radio images, which were used in an unsupervised search for objects with peculiar morphology in the ASKAP EMU pilot survey data. For all considered downstream tasks, we reported the model performance metrics and discussed the benefits brought by self-supervised pre-training, paving the way for building radio foundational models in the SKA era.

**Keywords:** Radio sources; radio source catalogs; astronomy image processing; deep learning; classification; outlier detection

(Received 29 April 2024; revised 9 July 2024; accepted 6 September 2024)

## 1. Introduction

Radio astronomy stands at the threshold of a transformative era, marked by the advent of large sky surveys carried out with instruments such as the Square Kilometre Array (SKA) (Dewdney et al., 2016) and its precursor telescopes. As the field enters this golden age, the immense volumes of observational data generated pose unprecedented challenges and opportunities. For example, the Evolutionary Map of the Universe (EMU) (Norris et al., 2011) of the Australian SKA Pathfinder (ASKAP, Johnston et al., 2008; Hotan et al., 2021) started in 2022 to survey ∼75% of the sky at 940 MHz with an angular resolution of $\sim15''$ and a noise rms of $\sim15\,\mu$Jy/beam. The EMU source cataloguing process will require an unprecedented degree of automation and knowledge extraction, as the expected number of detectable sources is ∼50 millions. So will be for other precursors and future SKA observations. The sheer scale and complexity of these datasets demand innovative approaches to shorten the time needed to deliver scientific results or groundbreaking discoveries.

In this context, machine learning (ML) emerges as a powerful tool for unlocking the full potential of radio astronomy data, offering solutions to complex tasks that are often beyond the reach of conventional methods in multiple areas, including source extraction, classification (e.g. morphological or astrophysical type) and discovery of anomalous/unexpected objects. Most existing contributions focused on galaxy morphology classification for extragalactic science cases employing either supervised learning (SL), e.g. with convolutional neural networks (CNNs) (Aniyan & Thorat 2017; Lukic et al., 2018; Wu et al., 2019; Lao et al., 2023) or Vision Transformers (ViTs) (Gupta et al., 2024), weakly-supervised learning (Gupta et al., 2023), semi-supervised learning (Slijepcevic et al., 2022), or unsupervised learning, e.g. Self-Organizing Maps (SOMs) (Galvin et al., 2020; Mostert et al., 2021; Gupta et al., 2022) or t-distributed stochastic neighbour embedding (Pennock et al., 2022).

Despite substantial progress, the full potential of supervised approaches often remains untapped due to the scarcity of large and high-quality annotated radio image datasets, crucial for training

**Corresponding author:** S. Riggi; Email: simone.riggi@gmail.com.

**Cite this article:** Riggi S, Cecconello T, Palazzo S, Hopkins AM, Gupta N, Bordiu C, Ingallinera A, Buemi C, Bufano F, Cavallaro F, Filipović MD, Leto P, Loru S, Ruggeri AC, Trigilio C, Umana G and Vitello F. (2024) Self-supervised contrastive learning of radio data for source detection, classification and peculiar object discovery. *Publications of the Astronomical Society of Australia* **41**, e085, 1–19. https://doi.org/10.1017/pasa.2024.84

very deep models. The human effort required to produce them is in fact unsustainable. Citizen science projects, launched within different precursor surveys on the Zooniverse platform[a,b] and building on the previous Radio Galaxy Zoo project (Banfield et al., 2015), will partially alleviate this need, at the cost of potentially introducing errors and biases in the cumulative dataset. As a result, existing labelled radio datasets are typically very limited in size, class-unbalanced, and adopt a diverse or ambiguous labelling schema, usually depending on the particular domain of application. Several applications produced so far for either radio source classification or source detection scopes, have therefore resorted to fine-tune models that were previously pre-trained on large annotated collections of non-astronomical data, such as the *ImageNet* (Deng et al., 2009) or *COCO* (Tsung-Yi et al., 2014) datasets, that may not well capture all distinctive features of radio observations. On the other hand, completely unsupervised approaches are not very effective when directly dealing with highly dimensional image data, typically requiring previous feature extraction and dimensionality reduction steps to be applied. Currently, employed methods based on SOMs typically enforce an apriori discrete static data organization that do not well support extension to new tasks. These limitations necessitate exploring alternative methodologies.

Representation learning (Bengio & Anal 2013), and in particular self-supervised learning (SSL) (Liu et al., 2023), has recently emerged as a promising avenue to address these issues, by directly learning (pretext task), without any supervision, a lower-dimensionality representation (i.e. the latent space) from large samples of unlabelled data. The resulting model and data representation can then be used for data inspection and generalized for various applications (downstream tasks), e.g. classification, object detection, etc, if a small subset of labelled data is available. Previous works in the radio domain are based on convolutional autoencoders (CAE) generative methods, which learns a latent space by minimizing a loss between input data and data reconstructed through an encoder-decoder network. For example, Ralph et al. 2019 developed a pipeline for unsupervised source morphology studies based on SOMs and k-mean clustering algorithm, employing CAEs to extract features from the Radio Galaxy Zoo (RGZ) (Banfield et al., 2015) images. Bordiu et al. (2023) employed CAEs to extract features from combined radio and infrared images of known Galactic supernova remnants (SNRs) to search for possible multiwavelength patterns.

Contrastive learning approaches, on the other hand, employ siamese or teacher-student network architectures, minimizing the similarity between augmented versions of the input data, eventually in contrast to negative samples. They were reported to obtain superior performance on natural images in classification tasks (e.g. rivalling supervised methods), quality of representation learnt, computation efficiency, and robustness to noise. Recently, Slijepcevic et al. (2024) trained BYOL (Grill et al., 2020) contrastive learning method over a sample of $\sim 10^5$ radio source RGZ images from the VLA FIRST survey (Becker et al., 1995). The resulting self-supervised model was then fine-tuned to classify FRI/FRII radio galaxies from the VLA FIRST survey, as listed in the *Mirabest* dataset (Porter & Scaife 2023). The analysis was repeated over a second dataset that include data from the MeerKAT MIGHTEE survey. Both analyses indicated an increase in classification accuracy (ranging from few percent to 8% for MIGHTEE) over the same model trained in a completely supervised way. Mohale & Lochner (2024) carried out a similar FRI/FRII classification analysis over the *Mirabest* dataset, using self-supervised models, previously pre-trained over the *ImageNet-1k* (natural images), RGZ (radio galaxy images), Galaxy Zoo DECaLS (optical galaxy images) datasets. Using a KNN classifier evaluation, they found that the model pre-trained on RGZ outperforms the others by a considerable margin (5% to 10% improvement in accuracy). Hossain et al. (2023) carried out the same analysis with both BYOL and SimCLR (Chen et al., 2020) self-supervised models but using Group Equivariant Convolutional Neural Network (G-CNN) backbones to make models invariant to different isometries (translation, rotation, mirror reflection). They pre-trained self-supervised models on the RGZ dataset and fine-tuned them on *Mirabest* dataset, obtaining FRI/FRII classification accuracies around 95%-97%[c], improving by ∼2% the fully supervised baseline.

With respect to previous studies, we focused more on SKA precursor data, training the SimCLR self-supervised model over large samples of unlabelled images, extracted from ASKAP and MeerKAT radio maps in two different modes: (1) “source-centered” mode, e.g. images centred and zoomed over catalogued sources (as in previous studies); (2) “blind” or “random” mode, e.g. images with arbitrary fixed size extracted from the entire map, without any source position awareness. The backbone component of the trained self-supervised models, a *ResNet18* neural network, was then evaluated and fine-tuned on labelled datasets to solve two representative radio source analysis tasks: radio source morphology classification and radio source instance segmentation. Additionally, the backbone model was used as a feature extractor for radio images, enabling an unsupervised search for radio objects with peculiar morphology based on the extracted data representation. Compared to previous studies, we assessed the trained models over larger labelled datasets, comprising thousands of annotated images from various radio surveys (VLA FIRST, ASKAP pilot, ATCA Scorpio), that were not previously used for self-supervised model pre-training. This study aims to quantify the benefits of self-supervision for the radio domain, providing ready-to-use foundational models that can be exploited in SKA precursor or other radio surveys as feature extractors for similar analysis or to tackle completely new tasks.

The paper is organized as follows. In Section 2 we describe the contrastive learning model considered, along with the training datasets, data pre-processing and training methodologies adopted. In Sections 3, 4, and 5 we studied how the trained self-supervised models perform in the three selected analysis scenarios, reporting performances achieved against benchmark supervised models. Finally, in Section 6 we summarize the obtained results and discuss future steps.

## 2. Self-supervised learning of radio data

### 2.1 Contrastive learning model

Figure 1 illustrates how self-supervised learning can be used for radio data analysis. Initially, a self-supervised framework (indicated by the red block) is trained on large samples of unlabelled image data. Subsequently, the resulting model backbone

[a] https://www.zooniverse.org/projects/chrismrp/radio-galaxy-zoo-lofar
[b] https://www.zooniverse.org/projects/hongming-tang/radio-galaxy-zoo-emu

[c] The observed metric differences between BYOL and SimCLR pre-trained models are not significant (below 1%) given the reported uncertainties.

https://doi.org/10.1017/pasa.2024.84 Published online by Cambridge University Press

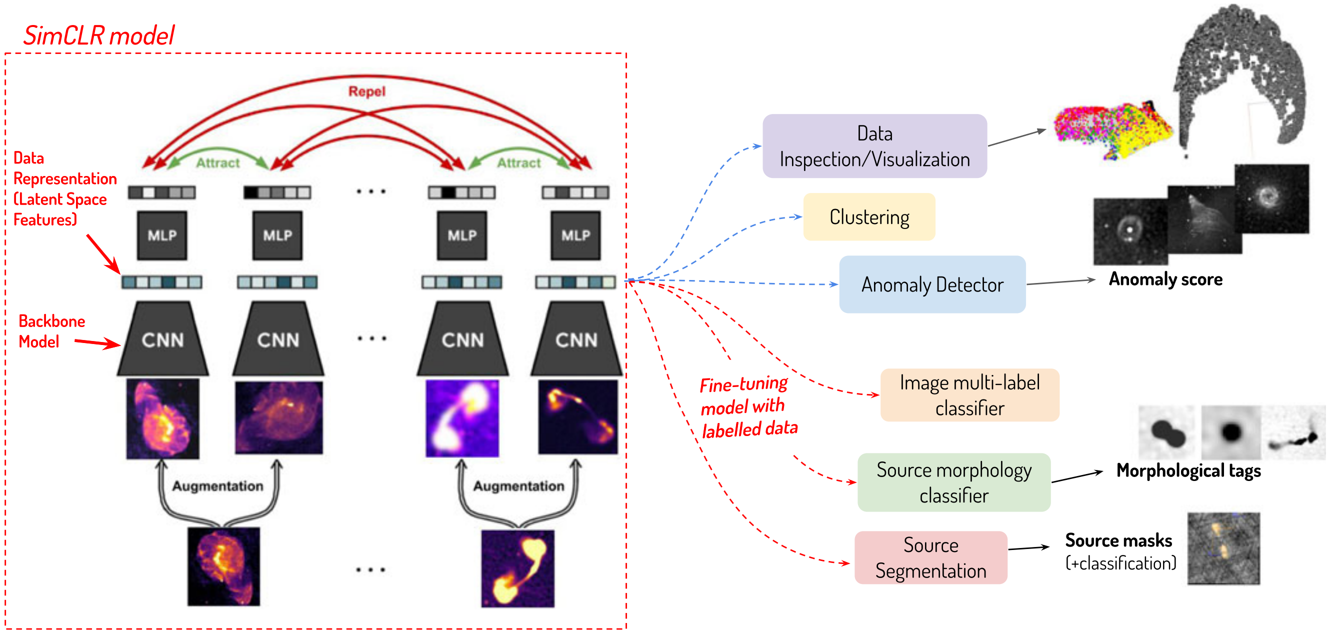

SimCLR model
Repel
Attract
Attract
Data Representation (Latent Space Features)
MLP
Backbone Model
CNN
Augmentation
Augmentation
Data Inspection/Visualization
Clustering
Anomaly Detector
Anomaly score
Fine-tuning model with labelled data
Image multi-label classifier
Source morphology classifier
Morphological tags
Source Segmentation
Source masks (+classification)

**Figure 1.** Schema of self-supervised learning for radio data analysis.

and data representation (or latent space vector) can be utilized for various downstream tasks, such as data inspection or anomaly detection, typically employing dimensionality reduction methods. Furthermore, the model can be applied to source detection and classification analysis using new labelled datasets. In this study, we used *SimCLR* as the self-supervised framework for our analysis.

SimCLR (Chen et al., 2020) is a simple yet widely used popular self-supervised learning framework. It learns data representations by maximizing the similarity between augmented views of the same input data (*positive examples*) relative to augmented views of different input data within the same training batch (*negative examples*). The architecture of SimCLR, depicted in Figure 1, consists of two main components: a base encoder network $f$, which is typically a *ResNet* network (He et al., 2016), and a small projection head network $g$, which is typically a Multi-Layer Perceptron (MLP) with one or two layers. Input images $\mathbf{x}_k$ ($k=1,...,N$) in a given batch sample of size $N$ are first processed to produce two augmented views (or positive pair) $\hat{\mathbf{x}}_{2k-1}$ and $\hat{\mathbf{x}}_{2k}$, by randomly applying multiple transformations from a specified transform set $\mathcal{T}$. The encoder network, also denoted as the *backbone model* throughout the paper, extracts representation vectors $\mathbf{h}_{2k-1}=f(\mathbf{x}_{2k-1})$ and $\mathbf{h}_{2k}=f(\mathbf{x}_{2k})$ from each augmented data pair. The projector network maps the representations to a space where a contrastive loss is applied, obtaining vectors $\mathbf{z}_{2k-1}=g(\mathbf{h}_{2k-1})$ and $\mathbf{z}_{2k}=g(\mathbf{h}_{2k})$. The contrastive loss $\mathcal{L}$, which is minimized during model training, is defined as:

$$\mathcal{L}=\frac{1}{2N}\sum_{k=1}^{N}\left[l_{2k-1,2k}+l_{2k,2k-1}\right] \tag{1}$$

$$l_{i,j}=-\log\frac{\exp(\mathrm{sim}(\mathbf{z}_i,\mathbf{z}_j)/\tau)}{\sum_{k=1}^{2N}\mathbb{1}_{k\neq i}\exp(\mathrm{sim}(\mathbf{z}_i,\mathbf{z}_k)/\tau)} \tag{2}$$

where $l_{i,j}$ is the normalized temperature-scaled cross entropy loss (NT-Xent), $\mathbb{1}_{k\neq i}=1$ if $k=i$ (equal to 0 otherwise), $\tau$ is a temperature parameter, and $\mathrm{sim}(\mathbf{z}_i,\mathbf{z}_j)$ is the pair-wise similarity between vectors $\mathbf{z}_i$ and $\mathbf{z}_j$, defined as:

$$\mathrm{sim}(\mathbf{z}_i,\mathbf{z}_j)=\frac{\mathbf{z}_i^T\mathbf{z}_j}{\parallel\mathbf{z}_i\parallel\parallel\mathbf{z}_j\parallel} \tag{3}$$

### 2.2 Datasets

We created the following unlabelled datasets for training SimCLR:

1. Two distinct datasets were generated using data from the SARAO MeerKAT Galactic Plane Survey (SMGPS) (Goedhart et al., 2024), which covers a large portion of the 1st, 3rd and 4th Galactic quadrants (l = 2°−61°, 251°−358°, $|b|<1.5°$) in the L-band (886–1 678 MHz). The survey has an angular resolution of 8″ and a noise rms of ∼10-20 $\mu$Jy/beam at 1.3 GHz:

   - `hulk_smgps`: A collection of 178,057 radio images, each of fixed size (256×256 pixels, equivalent to ∼6.4’×6.4’), extracted from SMGPS 1.28 GHz integrated intensity maps. This dataset was created by assuming a sliding window that traverses the entire surveyed area with a shift size equal to half the frame size, resulting in a 50% overlap among frames. The image size was chosen to be large enough to encompass the most extended radio galaxies that might be located in the cutout[d];
   - `banner_smgps`: A collection of 17 062 radio images extracted from SMGPS 1.3 GHz integrated maps, each centered around sources listed in the SMGPS

[d] Out of ∼5 800 catalogued sources that were labelled as candidate radio galaxies on the basis of their radio morphology, only one was found to have a size (7.4’) larger than the chosen image cutout (6.4’).

extended source catalogue (Bordiu et al., 2024). The size of the images varies across the dataset and is set to 1.5 times the size of the source bounding box. The radio sources in this dataset exhibit different morphologies, including single-island, multi-island, and diffuse sources.

2. Two distinct datasets were generated using data from the ASKAP EMU pilot survey (Norris et al., 2021a), which covered approximately 270 deg$^2$ of the Dark Energy Survey area, achieving an angular resolution of 11″ to 18″ and a noise rms of ∼30 $\mu$Jy/beam at 944 MHz:

   - `hulk_emupilot`: A collection of 55 773 radio images, each of fixed size (256×256 pixels, equivalent to ∼8.5’×8.5’), extracted from ASKAP EMU pilot 944 MHz integrated map. The images were extracted using a sliding frame that traversed the entire mosaic with a shift size equal to half the frame size, resulting in a 50% overlap among frames.
   - `banner_emupilot`: A collection of 10,414 radio images extracted from ASKAP EMU pilot 944 MHz integrated map, each centered around extended sources listed in the pilot source catalogue compiled by Gupta et al. (2024). The size of the images varies across the dataset and is set to 1.5 times the size of the source bounding box. The radio sources in this dataset exhibit different morphologies, including FR-I (∼6%), FR-II (∼54%), FR-x (∼14%), single-peak resolved (∼23%) radio galaxies. ∼3% of the sources present a rare morphology not fitting into the previously mentioned categories.

Datasets extracted in a blind mode (e.g. without any previous knowledge of the source location and morphology) can be constructed rapidly, potentially reaching substantial sizes (up to millions of images) when using future full-sky surveys. Without additional selection processes, these datasets tend to be largely unbalanced, predominantly comprising frames composed entirely of compact sources. The `hulk_smgps` dataset also comprises frames with large-scale diffuse emission, including background or portions of very extended sources located along the Galactic plane. For simplicity, we have labelled them as `hulk`. In contrast, “smarter” datasets centered on selected source positions typically have smaller sizes, requiring significant efforts (catalogue production and source type annotation) for construction. We have labelled them as `banner`. Indeed, one goal of this work is evaluating differences and benefits of both kind of datasets over different analysis tasks. Summary information for all produced datasets is reported in Table 1. In Figure 2 we display sample images from the `hulk_smgps` (top panels), `banner_smgps` (middle panels) and `banner_emupilot` (bottom panels) datasets.

**Table 1.** Summary information of datasets used for SimCLR model training. The number of images $n_{img}$ is reported in column (2). The image size $s_{img}$ is reported in column (3). $s_{img}$ is fixed for all images in the `hulk_smgp` and `hulk_emupilot` datasets, while $s_{img}$ is not fixed and depends on the source size $s_{\text{source}}$ (equivalent to the maximum source bounding box dimension) in the `banner_smgps` and `banner_emupilot` datasets. For these datasets, we report the average, minimum and maximum source sizes in columns (4), (5) and (6), respectively. Images from all datasets are eventually resized to a fixed size for model training and testing (see Section 2.3).

| | | | | $s_{\text{source}}$ | |
|---|---|---|---|---|---|
| *Dataset* | $n_{img}$ | $s_{img}$ | $\langle s_{\text{source}} \rangle$ | $s_{\text{source}}^{min}$ | $s_{\text{source}}^{max}$ |
| | | (pix) | (arcmin) | (arcsec) | (arcmin) |
| `hulk_smgps` | 178 057 | 256×256 | – | – | – |
| `hulk_emupilot` | 55 773 | | – | – | – |
| `banner_smgps` | 17 062 | 1.5×$s_{\text{source}}$ | 1.3 | 11.3 | 24.7 |
| `banner_emupilot` | 10 414 | | 1.2 | 18.8 | 7.8 |

## 2.3 Data pre-processing and augmentation

For the training and inference stages, we applied these pre-processing steps to input images:

- Grayscale images were converted to 3-channels. Each channel was processed differently from others, applying the following transformations:
  - *Channel 1*: sigma-clipping ($\sigma_{low} = 5$, $\sigma_{up} = 30$);
  - *Channel 2*: zscale transform (contrast = 0.25);
  - *Channel 3*: zscale transform (contrast = 0.4).
- Each channel was independently normalized to a [0,1] range using a *MinMax* transformation;
- Finally, each channel was resized to a 224×224 size in pixels.

A key aspect when training contrastive learning models is the choice of applied data augmentation steps to make the model invariant with respect to non-physical properties or to features not related to the radio sources. We applied the following augmenters to the data sequentially:

- *RandomCropResize*: randomly crop input images to size `crop_size` × image size, and resize data to the original size. `crop_size` was randomly varied in the range [0.8, 1.0];
- *ColorJitter*: apply a colour jitter transformation using all three image channels;
- *Flip*: random flip images either vertically or horizontally;
- *Rotate*: rotate images by either 90, 180 or 270 degrees;
- *Blur*: apply Gaussian blurring to images using a $\sigma$ parameter randomly varied in the range [1,3];
- *RandomThresholding*: threshold each channel separately using a per-channel percentile threshold randomly varied in the range [40,60].

The *RandomThresholding* augmenter was introduced to make the model less dependent on image background features. This stage was not applied when training over the `banner` datasets, as images in this dataset are already zoomed on radio sources, and thus the background would likely not be estimated correctly. Furthermore, not all augmenters are applied to every image in the training dataset. In Table 2 we provide a summary of augmenter steps used in the pre-processing pipeline with their parameters, including the probability with which each data transform is applied to images. With respect to Chen et al. (2020), we reduced the fraction of random cropping allowed to avoid cutting out relevant details of extended sources from the resulting image.

https://doi.org/10.1017/pasa.2024.84 Published online by Cambridge University Press

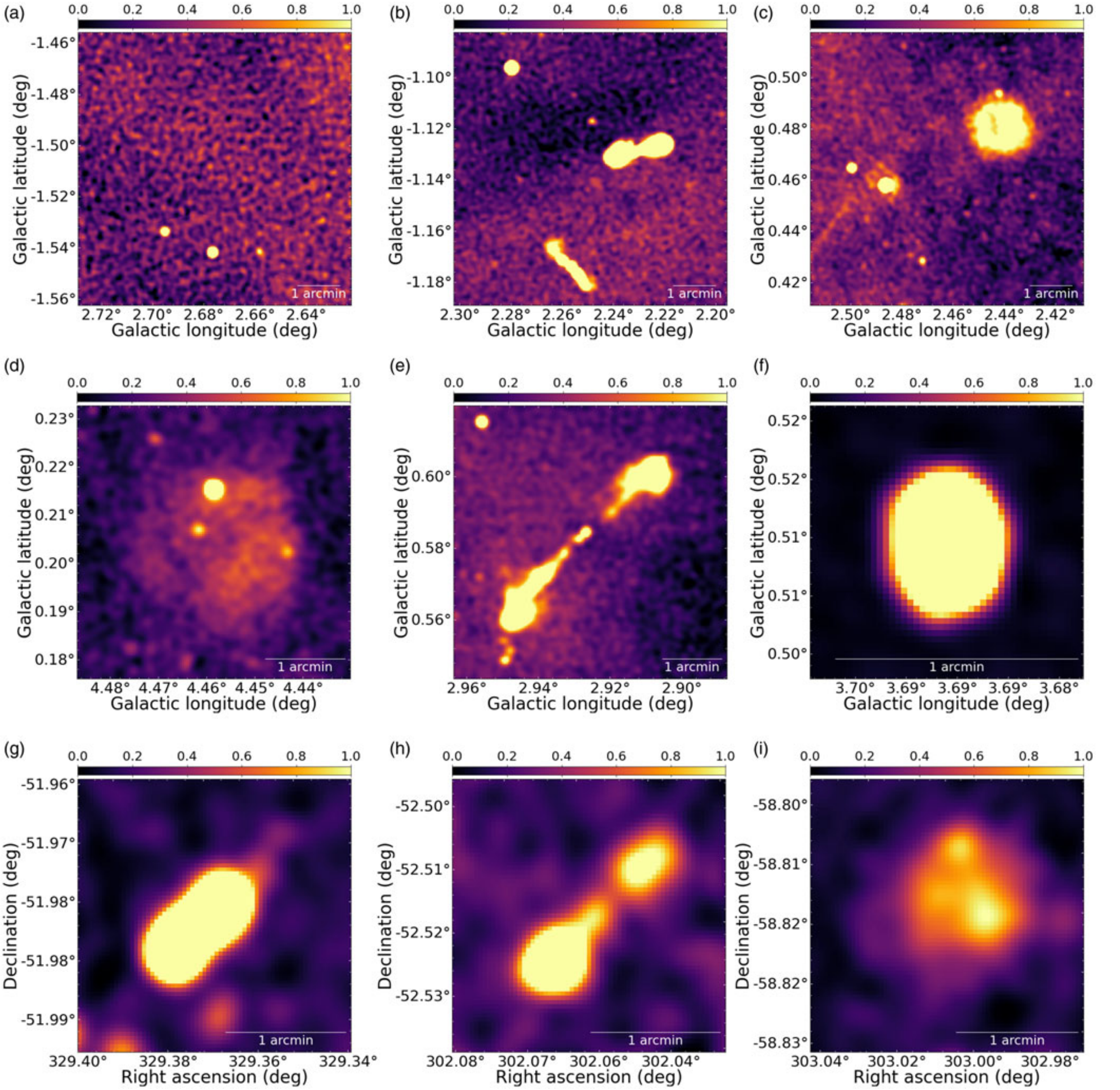


**Figure 2.** Representative examples of images from the `hulk_smgps` (top panels), `banner_smgps` (middle panels) and `hulk_emupilot` (bottom panels) datasets. A zscale transform was applied to all images for visualization scopes. Top panels: sample images containing only compact sources (Figure 2(a)), or multiple extended sources (Figures 2(b) and 2(c)). Middle panels: sample source with diffuse morphology (Figure 2(d)), a multi-component extended source exhibiting typical radio galaxy morphology (Figure 2(e)), a single-component extended source with a roundish morphology (Figure 2(f)). Bottom panels: sample sources with FR-I (Figure 2(g)), FR-II (Figure 2(h)) and peculiar (Figure 2(i)) classification.

### 2.4 Model training

We trained a SimCLR model on each of the four datasets described in Section 2.2, using the hyperparameters listed in Table 3. We will refer to them using their training dataset name: `hulk_smgps`, `banner_smgps`, `hulk_emupilot`, and `banner_emupilot`. A fourth model, referred to as `smart_hulk_smgps` hereafter, was trained in two steps, first on the `hulk_smgps` dataset and then on the `banner_smgps` dataset. The final model weights from the first step were used as initialization for the second step. For all models, we used a *ResNet18* (He et al., 2016) encoder and a 2-layer projector with 256 and 128 neurons, respectively.

Following Chen et al. (2020), all training runs began with a linear warm-up phase lasting 10 epochs, after which we switched to a cosine learning rate decay strategy. In total, we trained models

**Table 2.** List of augmentations used in SimCLR model training. In column (2) we reported the transform parameter values. In column (3) we reported the probability used to apply the transform in the augmentation pipeline, e.g. 1.0 means the transform is always applied to all input images.

| Augmentation | Parameters | Probability |
|---|---|---|
| *RandomCropResize* | `crop_size` = [0.8,1.0] | 1.0 |
| *ColorJitter* | `brightness` = 0.8 | 0.8 |
| | `contrast` = 0.8 | |
| | `saturation` = 0.8 | |
| | `hue` = 0.2 | |
| | `strength` = 0.5 | |
| *HorizontalFlip* | – | 0.33 |
| *VerticalFlip* | – | 0.33 |
| *Rotate* | `angle` = {90,180,270} | 0.5 |
| *Blur* | `sigma` = [1,3] | 0.1 |
| *RandomThresholding* | `percentileThr` = [40,60] | 0.5 |

**Table 3.** List of hyperparameters used in SimCLR model training.

| Hyperparameter | Value |
|---|---|
| Encoder model | ResNet18 |
| Projector model size | 2 layers (256-128) |
| Optimizer | Adam |
| Batch size | 128 |
| Learning Scheduler | Linear warmup + Cosine decay |
| Warmup epochs | 10 |
| Learning rate (warmup target) | 0.1 |
| Epochs | 100 (`hulk_smgps` datasets) |
| | 500 (`banner_smgps` dataset) |
| | 100 (`hulk_emupilot` dataset) |
| | 500 (`banner_emupilot` dataset) |

for 500 epochs on the `banner_smgps` and `banner_emupilot` datasets. A smaller total number of epochs (100) was used when training over the larger `hulk_smgps` and `hulk_emupilot` datasets to reduce computational costs.

Training runs were performed on three different computing server nodes, each equipped with a GPU device:

- Node A: 48 cores (Intel Xeon Gold 6248R CPU, 3.00 GHz), 512 GB RAM, NVIDIA Quadro RTX 6000 (24 GB)
- Node B: 24 cores (Intel Xeon Silver 4410Y, 2.00 GHz), 256 GB RAM, NVIDIA A30 (24 GB)
- Node C: 36 cores (Intel Xeon CPU E5-2697 v4, 2.30 GHz), 128 GB RAM, NVIDIA Tesla V100 (16 GB)

Typical training times over the `hulk_smgps` dataset are of the order of ∼6.7 hours/epoch on nodes A/B, and ∼12.5 hours/epoch on node C.

### 2.5 *Evaluation of downstream tasks*

In the following sections, the trained self-supervised models and corresponding data representation will be evaluated on radio source classification (Section 3) and detection (Section 4) tasks using supervised CNN classifiers trained on labelled datasets. Furthermore, in Section 5 we will use the self-supervised features to classify radio images according to the morphology of hosted sources in a supervised way and according to their peculiarity degree in a completely unsupervised way. To estimate the performances achieved in these downstream tasks, we will consistently use these widely adopted metrics in multi-class problems:

- *Recall* ($\mathcal{R}$): Fraction of sources (images) of a given class that were correctly classified by the model out of all sources (images) labelled in that class, computed as:

$$\mathcal{R} = \frac{TP}{TP + FN}$$

- *Precision* ($\mathcal{P}$): Fraction of sources (images) correctly classified as belonging to a specific class, out of all sources (images) the model predicted to belong to that class, computed as:

$$\mathcal{P} = \frac{TP}{TP + FP}$$

- *Contamination* ($\mathcal{C}$): Fraction of sources (images) of a given class incorrectly classified as belonging to a specific class, out of all sources (images) the model predicted to belong to that class, computed as:

$$\mathcal{C} = \frac{FP}{TP + FP} = 1 - \mathcal{P}$$

- *F1-score*: the harmonic mean of precision and recall:

$$\text{F1-score} = 2 \times \frac{\mathcal{P} \times \mathcal{R}}{\mathcal{P} + \mathcal{R}} \qquad (4)$$

where $TP$, $FN$, $FP$ are the number of true positive, false negative and false positive instances, respectively.

## 3. Task I: Classification of radio source morphology

In this section, we quantitatively evaluate the learned self-supervised representation on a source morphology classification problem.

Morphological classification plays a pivotal role in radio astronomy, enabling scientists to gain insights into the underlying source nature from the observed shape and structures. The majority of existing works in the radio image domain are targeted for extragalactic science objectives, focusing on classification of radio galaxies (see for example Aniyan & Thorat 2017, Ma et al., 2019, or Ndung'u et al., 2023 for a recent review) in different morphological classes: `compact`, `FR-I`, `FR-II`, `bent-tailed` (including WAT[e] and NAT[f] population), `irregular` (including, for example, X-shaped or ring-like radio galaxies).

Morphological classification is also an important post-detection stage to filter extracted sources by general morphology (e.g. compact vs extended sources) for specialized source property measurements or other advanced classification analysis. In this context, the adopted source labelling scheme is rather general-purpose and domain-agnostic, suited to be eventually refined afterwards. For example, typical used labels are

[e] Wide-angle tail

[f] narrow-angle tail

https://doi.org/10.1017/pasa.2024.84 Published online by Cambridge University Press

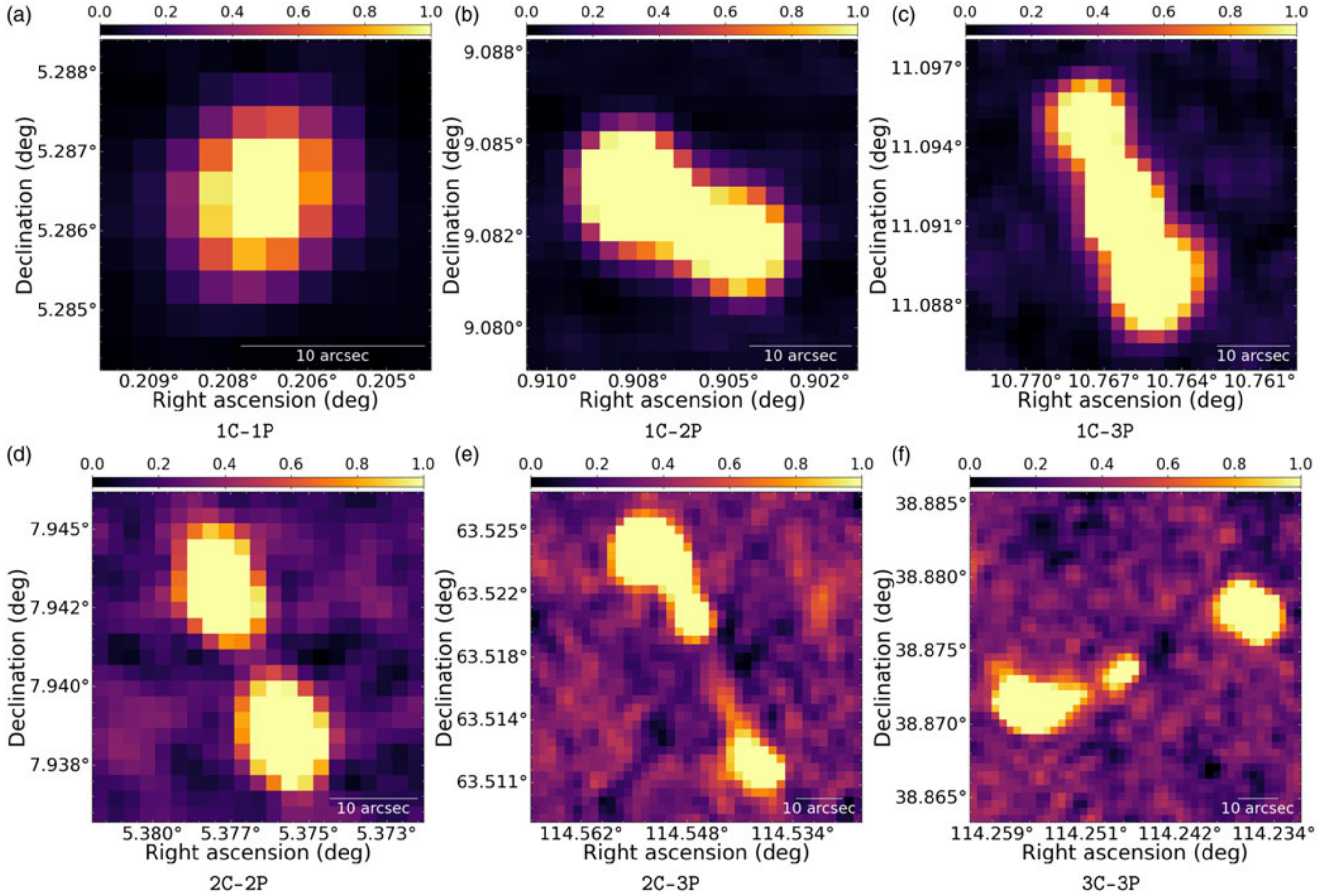


**Figure 3.** Sample images from the RGZ dataset with representative sources of different morphological classes (reported below each frame). A zscale transform was applied to all images for visualization scopes.

POINT-LIKE, RESOLVED, COMPACT, EXTENDED or labels that contain information about the number of radio components present in the extracted source (as in Wu et al., 2019).

The analysis carried out in this section falls into the second use-case scenario. This choice is mostly driven by existing datasets. Available annotated datasets for radio galaxy classification (mostly based on VLA FIRST survey data) are, in fact, rather limited in size (e.g. typically <100-200 images per class, <2000 images overall) and would currently prevent us from obtaining a robust evaluation of our self-supervised models over multiple test set realizations. For example, the *Mirabest* dataset (Porter & Scaife 2023) contains 1256 source images of balanced FR-I/FR-II radio galaxy classes, out of which 833 images constitute the "Confident" sample, and the rest (423 images) the "Uncertain" sample. On this dataset, Slijepcevic et al. (2024) reported an improvement of ∼3-4% in classification performance of a self-supervised pre-trained model with respect to a fully supervised model trained from scratch on the "Confident" sample (or on a portion of it). Classification metrics were, however, estimated on the "Uncertain" sample, and hence the observed enhancement is due to less than 20 sources. We, therefore, opted for this work to use a larger dataset (roughly by one order of magnitude) and perform a similar analysis once a larger dataset is assembled within the ASKAP EMU survey.

### 3.1 Dataset

For this analysis, we considered data from the Radio Galaxy Zoo (RGZ) project[g] (Banfield et al., 2015). This includes radio images of size 3'× 3' from the VLA Faint Images of the Radio Sky at Twenty cm (FIRST) survey (1.4 GHz, angular resolution ∼5″) (Becker et al., 1995). Radio sources found in these images were labelled into multiple morphological classes, on the basis of the observed number of components (C) and peaks (P) (see Wu et al. 2019 for more details on the classification schema). Angular size is also available for each source.

In this analysis, we extracted 82 084 image cutouts around radio sources that have been classified in the RGZ Data Release 1 (DR1) with a consensus level ≥0.6 in the following classes: 1C-1P (55.0%), 1C-2P (20.9%), 1C-3P (1.9%), 2C-2P (17.6%), 2C-3P (2.0%), 3C-3P (2.5%). We assumed a cutout size equal to 1.5 × the source angular size, as listed in the RGZ catalogue. A representative image of each source morphological category is shown in Figure 3.

[g] The RGZ project is a crowdsourced science project where both scientists and citizens can classify radio galaxies and their host galaxies from radio and infrared (WISE survey, Wright et al. 2010) images presented to users in a web interface

https://doi.org/10.1017/pasa.2024.84 Published online by Cambridge University Press

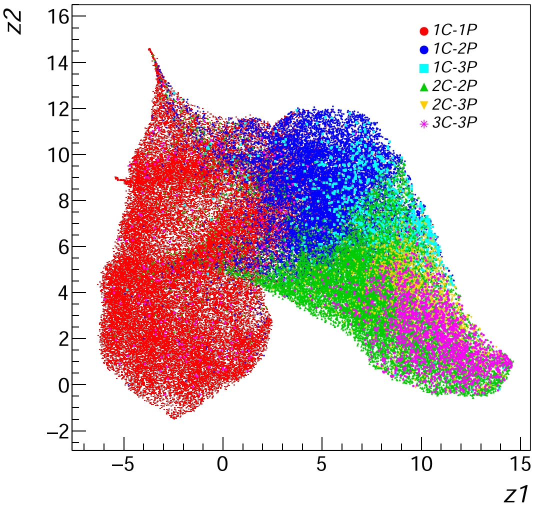


**Figure 4.** 2D UMAP projection of the data representation vector (size = 512) produced by the trained `smart_hulk_smgps` model on the RGZ dataset.

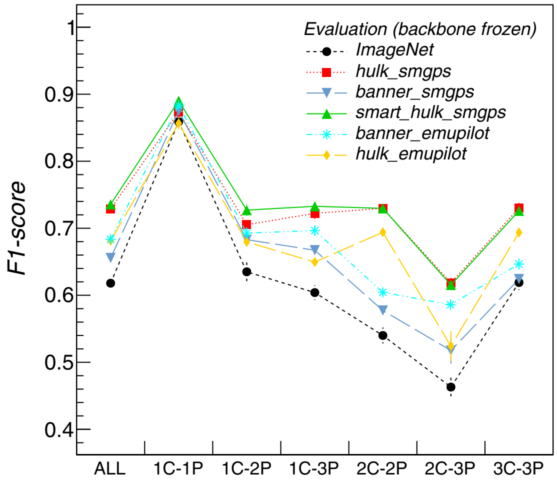


**Figure 5.** Classification F1-scores obtained for different classes and for all classes cumulatively over RGZ test sets with different pre-trained and frozen backbone models: `hulk_smgps` (red squares), `banner_smgps` (blue inverted triangles), `smart_hulk_smgps` (green triangles), `hulk_emupilot` (orange diamonds), `banner_emupilot` (cyan asterisks), `ImageNet` (black dots). The reported values and errors are the F1-score mean and mean error computed over five test sets.

As the full dataset is largely unbalanced towards sources of class morphology 1C-1P, we randomly created $N_{sets} = 5$ balanced training and test sets having 1000 and 600 images per class, respectively. Both training and test set images were pre-processed as described in Section 2.3 for the SimCLR model training.

### 3.2 Evaluation of self-supervised representation

In Figure 4 we present a two-dimensional projection, obtained with the *Uniform Manifold Approximation and Projection* (UMAP) (McInnes et al., 2018) dimensionality reduction algorithm, of the representation vector (original size equal to 512) produced by the trained `smart_hulk_smgps` model on the RGZ dataset. As can be observed, the self-supervised model effectively groups sources of different morphological class in distinct areas of the latent space. No isolated clusters are discernible in the projected two-dimensional UMAP feature space, as well as in a PCA scatter plot of top-2 features (not shown here). Nevertheless, these or similar diagnostic plots, can be useful for potentially identifying possible image mislabeling in the dataset, e.g. sources that fall within a region that is predominantly populated by other classes.

We carried out a classification analysis using a CNN classifier with a standard architecture: a *ResNet18* backbone model (as in the SimCLR model) followed by a classification head. The latter consists of a single layer followed by a softmax activation, representing the predicted probability distribution over the set of classes. To evaluate the quality of the self-supervised representation, we froze the backbone model, setting and fixing its weights to those obtained in the trained SimCLR models, and trained only the classification head on RGZ training datasets for a limited number of epochs (30). We considered only rotation and flipping transformations as augmentation steps during the training.

In Figure 5 we report the classification F1-scores obtained on the test set by different self-supervised pre-trained models: `hulk_smgps` (red squares), `banner_smgps` (blue inverted triangles), `smart_hulk_smgps` (green triangles), `hulk_emupilot` (orange diamonds), `banner_emupilot` (cyan asterisks). The reported values and their errors are respectively the F1-score mean and mean error computed over the available test sets. These metrics were compared against those obtained with a baseline model pre-trained on the `ImageNet-1k` dataset[h] Deng et al., 2009) (trained on non-radio data), shown with black dots in Figure 5. We found that self-supervised pre-trained models reach approximately 7–12% better overall scores with respect to the baseline, due to the higher quality features obtained on complex and extended sources, which are not as well represented in the *ImageNet* dataset. Another valuable indication is that the two-step pre-training approach done for the `smart_hulk_smgps` model training provide better results compared to training over random or source-centred images alone. The improvement is, however, not very significant with respect to the `hulk_smgps`, likely due to both the limited size of the `banner_smgps` dataset and the absence of Galactic-like diffuse and extended sources in the RGZ dataset. By construction, we expect that the `banner_smgps` model should be more specialized for this kind of source morphologies. This will be tested in a future analysis once we finalize a new test dataset with diffuse sources taken from ASKAP EMU observations.

In Figure 6 we report the confusion matrix obtained over the RGZ test sample with a `hulk_smgps` pre-trained and frozen backbone model. The obtained misclassification rates suggest that a considerable fraction (10% to 20%) of sources, particularly those with two or three components, may be hard to be correctly distinguished from other classes. After a visual inspection of the misclassified sources, we found that in some cases the misclassification is rather due to dataset mislabelling, e.g. the ground truth label present in the dataset is not correct and the model is indeed predicting the expected class. Some examples are reported in Figure A.1. Future analysis should therefore take into consideration a revision of the RGZ dataset annotation.

[h] When mentioning the `ImageNet` dataset throughout the paper, we refer to the `ImageNet-1k` version.

https://doi.org/10.1017/pasa.2024.84 Published online by Cambridge University Press

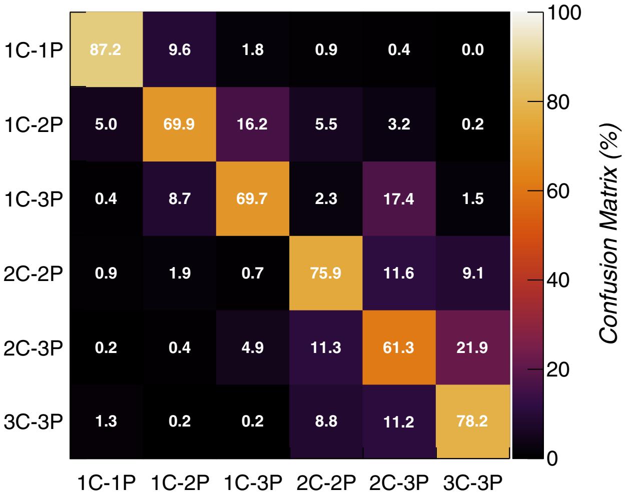


**Figure 6.** Confusion matrix of the source morphology classifier (trained with `smart_hulk_smgps` pre-trained and frozen backbone model) obtained over the RGZ test set.

### 3.3 Model fine-tuning

We fine-tuned the source classifier by unfreezing backbone model layers (e.g. training them along with the classification head) and compared the accuracies reached by two models: one initialized with random weights (e.g. training from *scratch*), and the other with backbone model weights initialized to the `smart_hulk_smgps` backbone model weights (best performing model found in Section 3.2). We compared the results of both models when trained on the full training sets and when trained on smaller training sets, obtained by gradually removing labelled data randomly from the original set. In all cases, models were trained for 150 epochs. The test sets were kept unchanged to compute the classification accuracies. This was done to study how the model performs in the recurring scenario in which the amount of labelled data is significantly limited. We reported the results in Figure 7. As can be seen, the fully supervised model (trained from scratch) becomes almost untrainable, providing poor classification metrics, in the small number of labels regime. This occurs for the RGZ dataset below a fraction of approximately 10% of the original training dataset, corresponding to about 600 images (e.g. ~100 images per class). On the other hand, self-supervised pre-training enables to fine-tune the model even with few labels, achieving considerably better metrics (>20%). Above the 10% label fraction threshold, the fully supervised model achieved slightly better scores, highlighting that no significant performance benefits are obtained from the pre-training process, at least with the model and dataset sample sizes available for this work. This result is qualitatively on par with the results of the transfer learning analysis carried out by Chen et al. (2020) (Section B.8.2) on 12 natural image datasets (Food, CIFAR10, CIFAR100,) with *ResNet50* architectures of different widths (x1, x2, x4). The authors compared the fine-tuning classification accuracies reached by SimCLR against a supervised model baseline. With wider networks (*ResNet50* x4), the self-supervised model was found to outperform the supervised one in 7 datasets (Chen et al. 2020, Table 8). The opposite was however observed with the narrower *ResNet50*, where the supervised baseline best performed in 10 datasets (Chen et al. 2020, Table B5) out of 12. Our analysis, carried out with an even smaller network (*ResNet18*), may well fall into the latter case. In either cases, the observed accuracy differences are smaller than 1% for most datasets.

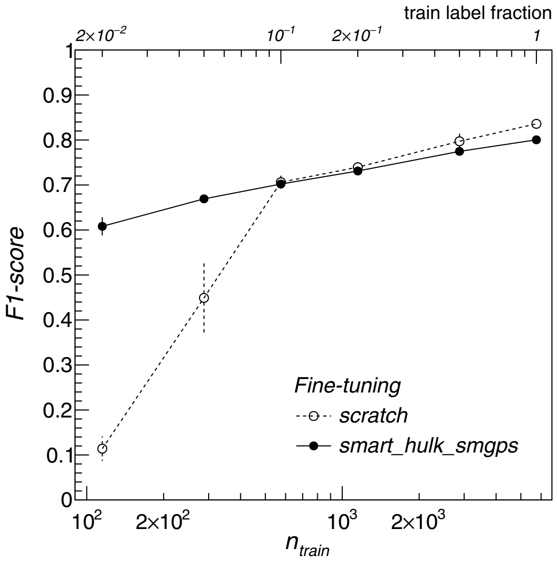


**Figure 7.** Classification F1-scores obtained (for all classes cumulatively) over RGZ test sets as a function of the number of images $n_{train}$ in the training set with two alternative models: one trained from scratch (open black dots), the other trained with backbone model weights initialized to `smart_hulk_smgps` weights (filled black dots). The upper x-axis indicates the fraction of the full training set considered in each training run.

## 4. Task II: Radio source detection

In this section, we quantitatively evaluate the learned self-supervised representation on an instance segmentation problem, specifically the detection of radio sources with various morphologies.

Algorithms used in traditional radio source finders are not well-suited for detecting extended radio sources with diffuse edges, and they are unable to detect extended sources that are composed of multiple disjoint regions. To address this limitation, new source finders (Wu et al., 2019; Mostert et al., 2022; Zhang et al., 2022; Yu et al., 2022; Riggi et al., 2023; Lao et al., 2023; Gupta et al., 2024; Cornu et al., 2024) based on deep neural networks and object detection frameworks have been developed and trained on either simulated or real radio data. Core components of these models are deep CNN backbones and transformer architectures, both of which have millions of parameters that need to be optimized during training. Although these models offer a substantial advancement in extended radio galaxy detection, their performance is limited by the small size (few thousand images) and the imbalance of objects in the available radio training datasets. Additionally, there is a potential performance drop (up to 10% in Riggi et al., 2023) when transferring a trained model to data produced by a different survey or telescope, especially if the new data has a better angular resolution (Tang et al., 2019). To improve the training stage, it is a common practice to use models pre-trained on much larger annotated samples of non-astronomical images, such as the *ImageNet-1k* (Deng et al., 2009, ~1.28 million images) or the *COCO* (Tsung-Yi et al. 2014, 328 000 images) datasets. In this scenario, it is worth exploring whether foundational models built with self-supervised methods on unlabelled radio data can

https://doi.org/10.1017/pasa.2024.84 Published online by Cambridge University Press

offer performance benefits over non-radio foundational models, especially with small datasets.

### 4.1 caesar-mrcnn source detector

For this analysis, we used the *caesar-mrcnn* source detector (Riggi et al., 2023), based on the Mask R-CNN object detection framework (He et al., 2017), to extract source segmentation masks and predicted class labels from input radio images. With respect to our original work (Riggi et al., 2023), we have upgraded the software to TensorFlow 2.x, producing a new refactored version[i] with an improved data pre-processing pipeline and support for additional backbone models.

In this context, we would like to make a brief preamble and clarify the motivations that guided the development of the *caesar-mrcnn* source detector, as these were either misinterpreted or inaccurately presented in other works. Additionally, we aim to address certain conceptual aspects that we realize are often source of confusion within this field.

It is essential to recognize that source detection (or extraction), classification and source characterization (or measurement) represent distinct conceptual stages. A source detector, to be defined as such, should focus solely on extracting source bounding boxes or, preferably, pixel masks, which are the inputs required for the source measurement or classification stages. The source measurement step, on the other hand, is responsible for estimating source properties such as position, flux density, and shape from the outputs of the source detection. Strictly speaking, this step is not required in a source detector, as assumed in Lao et al. (2023). From a methodological standpoint, it is advisable to avoid conflating these stages. This may allow addressing numerous use cases simultaneously, but it can also be counterproductive, leading, for example, to design compromises and overly complex models with multiple loss components to be balanced during training. The resulting models likely have a higher chance of underperforming on both tasks (detection or characterization) with respect to models that are designed and optimised for a specific task. For this reason, source characterization metrics should be independently evaluated and not mixed with the detection metrics, as required, for example, in the SKA Data Challenge 1 (Bonaldi et al., 2021) scoring function. When we designed the *caesar-mrcnn* source detector, we deliberately did not provide a source characterization stage. As we already implemented source measurement functions in the *caesar* source finder, we rather aim to interface both codes and, at best, add new developments for improvements in specific areas, such as low S/N source characterization and source deblending, as discussed in Boyce et al. (2023).

In recent ML-based source extractors, source classification was typically performed alongside the detection step, often to classify extracted sources into compact and extended classes of radio galaxies (FR-I, FR-II, etc.). We aimed for our source detector to be general-purpose, portable, and not tied to a specific radioastronomical domain. Therefore, in our view, the detection step should, at a minimum, classify between real and spurious sources, or, at most, between domain-agnostic morphological classes. More refined or domain-specific classification schemes can be more effectively applied afterwards in specialized classifiers working on source-centered images obtained from the detection step. These considerations were the rationale behind the general class labeling scheme adopted in *caesar-mrcnn* (briefly reported in the following Section).

### 4.2 Dataset

To train and test *caesar-mrcnn*, we considered the dataset described in Riggi et al., 2023), which contains 12 774 annotated radio images from different surveys, including VLA FIRST, ATCA Scorpio (Umana et al., 2015), and ASKAP-EMU Scorpio (Umana et al., 2021). The annotation data consist of pixel segmentation masks and classification labels for a total of 38 342 objects (both real and spurious sources) present in the dataset images. Five object classes were defined:

- `SPURIOUS`: imaging artefacts around bright sources, having a ring-like or elongated compact morphology;
- `COMPACT`: single-island isolated point- or slightly resolved compact radio sources with one or more blended components, each with morphology similar to the synthesized beam shape;
- `EXTENDED`: radio sources with a single-island extended morphology, with one or more blended components, some morphologically different from the synthesized beam shape;
- `EXTENDED-MULTISLAND`: radio sources with an extended morphology, consisting of more than one island, each eventually containing one or more blended components, having a point-like or an extended morphology;
- `FLAGGED`: poorly-imaged single-island radio sources, highly contaminated by nearby imaging artefacts.

For more details on the dataset labelling schema and rationale, we refer the reader to the original work. We also define a generic class label `SOURCE` for analysis purposes, including real and non-flagged sources, i.e. object instances of class `COMPACT`, `EXTENDED`, or `EXTENDED-MULTISLAND`. Though it is planned, the dataset does not presently contain images and annotation data for Galactic diffuse objects. Indeed, none of existing ML-based finders have been trained to detect diffuse sources other than radio galaxy diffuse structures (e.g. lobe components). The latter are the only diffuse structures present in our dataset, but we never noted to obtain poor detection performances on them, as reported in Ndung'u et al. (2023).

In Figure 8 we present sample images from the dataset, including representative sources for each class. Given that the current dataset is significantly skewed towards compact sources (comprising approximately 80% of the annotated objects), we created five re-balanced training samples, each containing 3245 images, with the following class distributions: `SPURIOUS` (1464 objects, 14.4%), `COMPACT` (5457 objects, 53.6%), `EXTENDED` (2042 objects, 20.1%), `EXTENDED-MULTISLAND` (1047, 10.3%), `FLAGGED` (169 objects, 1.7%). The remaining data was reserved to create five test samples, each containing 5110 images, with the following class distributions: `SPURIOUS` (1022 objects, 6.6%), `COMPACT` (12.346 objects, 80.0%), `EXTENDED` (1307 objects, 8.5%), `EXTENDED-MULTISLAND` (636, 4.1%), `FLAGGED` (122 objects, 0.8%).

### 4.3 Evaluation of self-supervised representation

To assess the effectiveness of the self-supervised representation, we followed the procedure outlined in Section 3.2. We froze

[i] https://github.com/SKA-INAF/caesar-mrcnn-tf2

https://doi.org/10.1017/pasa.2024.84 Published online by Cambridge University Press

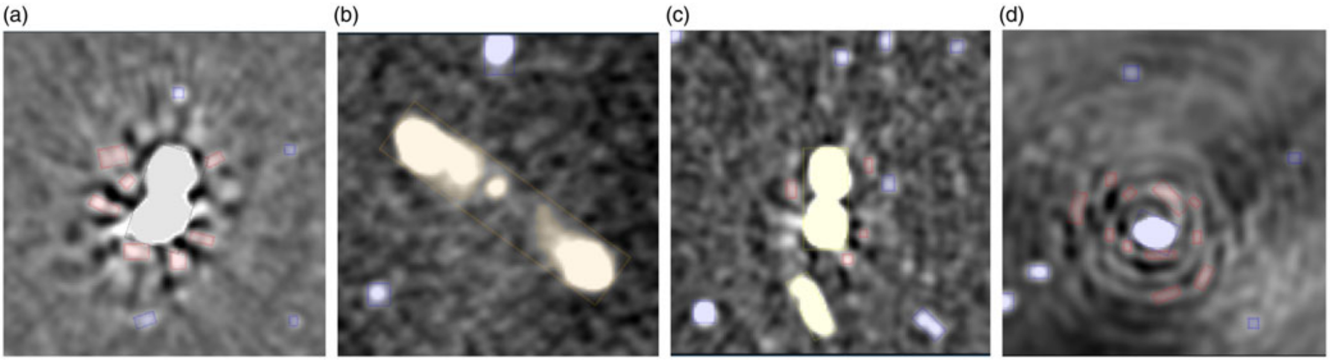


**Figure 8.** Sample images (taken from Riggi 2023) from the dataset used for *caesar-mrcnn* training/testing, including objects of different classes: a `FLAGGED` object (Figure 8(a), in gray), `COMPACT` objects (in blue), a `MULTI-ISLAND` object (Figure 8(b), in orange), `EXTENDED` objects (Figure 8(c), in yellow), `SPURIOUS` objects (Figure 8(d), in red).

the Mask R-CNN *ResNet18* backbone model, setting and keeping its weights fixed to those obtained in the trained SimCLR models, and trained the remaining components (region proposal network, classification and bounding box regression head, mask prediction head) on multiple training datasets for a set number of epochs (250). The parameters of Mask R-CNN were configured to match the values optimized in our previous work (refer to Riggi 2023, Table A1). We applied the same pre-processing transformations used for training the self-supervised models (as detailed in Section 2.3). During training, we applied three distinct image augmentations: rotation, horizontal/vertical flipping, and zscale transformation with random contrast in the range of 0.25 to 0.4.

The performance of source detection was evaluated on the test sets using the metrics[j] defined in Section 2.5 and the following detection/classification criteria:

1. Object detection score threshold equal to 0.5;
2. Intersection-over-Union (IoU) match threshold between detected and ground truth object masks equal to 0.6;
3. Object classified in the `SOURCE` class group.

The above metrics were computed for each class label and reported in Figure 9 for different models trained with frozen self-supervised backbone model weights: `hulk_smgps` (red squares), `banner_smgps` (blue triangles), `smart_hulk_smgps` (green triangles), `hulk_emupilot` (orange diamonds). Metrics obtained with frozen *ImageNet* weights are shown with black dots. The performance boost obtained with self-supervised models is significant, around 15%-20% for most classes, and even larger for multi-island sources and imaging artefacts. This is somehow expected, given that these structures are not present in the *ImageNet* dataset. Overall, for the source class group we are interested in, we did not notice significant differences among trained self-supervised models, after taking into account the run-to-run statistical uncertainties on the obtained metrics. We will therefore consider a representative model (`hulk_smgps`) in the following fine-tuning analysis.

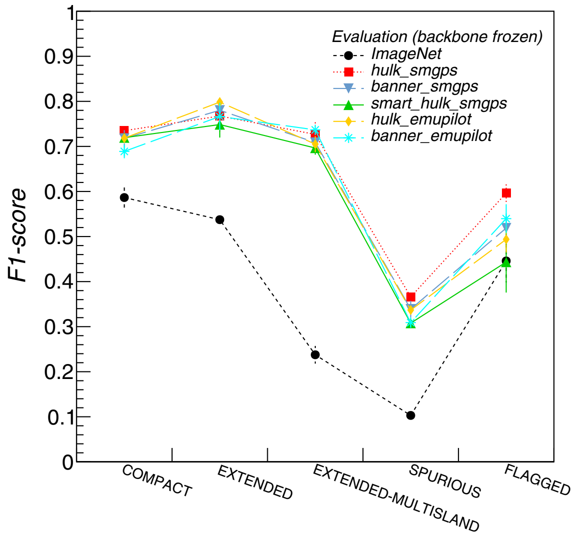


**Figure 9.** Mask R-CNN object detection F1-score metric obtained for different object classes over multiple test sets with different pre-trained and frozen backbone models: `hulk_smgps` (red squares), `banner_smgps` (blue iverted triangles), `smart_hulk_smgps` (green triangles), `hulk_emupilot` (orange diamonds), `banner_emupilot` (cyan asterisks), `ImageNet` (black dots). The reported values and errors are the means and mean errors computed over 5 test sets.

### 4.4 Model fine-tuning

We fine-tuned the Mask R-CNN model using random initialization weights (training from *scratch*) and weights initialized to `hulk_smgps` self-supervised model. We computed the object detection metrics over the source class group as a function of the training sample size, following the same approach discussed in Section 3.3. Results are reported in Figure 10. Black filled dots are the F1-scores obtained with the pre-trained `hulk_smgps` model, while open black dots are those found when training from scratch. In this case, we did not observe a significant benefit from using self-supervised pre-training compared to the source classification task studied in Section 3. The improvement in performance in the low label regime (<10% of the original training sample size) is, in fact, of the order of a few percent. This behaviour highlights that other Mask R-CNN components likely play a major role in the

[j] In astronomical source catalogue works, the recall/precision metrics are often referred to as completeness/reliability.

https://doi.org/10.1017/pasa.2024.84 Published online by Cambridge University Press

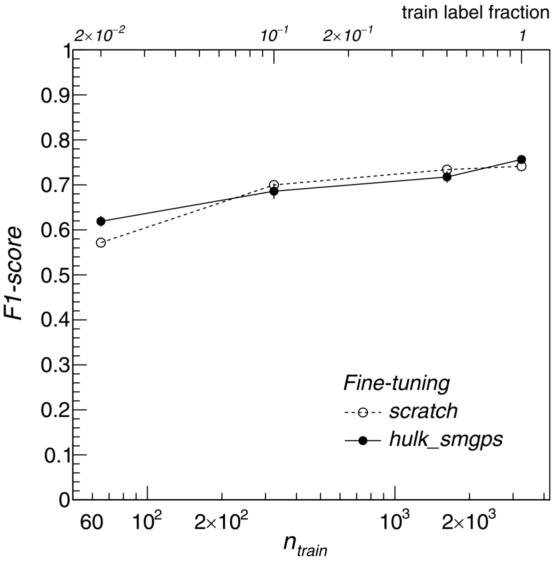


**Figure 10.** Mask R-CNN object detection F1-score metric obtained over the `SOURCE` class over multiple test sets as a function of the number of images $n_{train}$ in the training set with two alternative models: one trained from scratch (open markers), the other trained with backbone model weights initialized to `hulk_smgps` weights (filled markers). The upper x-axis indicates the fraction of the full training set considered in each training run.

overall model detection performance with respect to the backbone model.

## 5. Task III: Search for peculiar objects

In this section, we quantitatively evaluated the learned self-supervised representations in an anomaly detection problem, i.e. employing them for an unsupervised search of radio objects with peculiar morphologies.

Next-generation radio surveys carried out with SKA precursor telescopes are already generating a huge amount of data. Serendipitous discoveries were already reported and obtained by visual inspection of the observed maps. For instance, Norris et al. (2021b) and Koribalski et al. (2021) discovered a class of diffuse objects with a roundish shape, dubbed *Odd Radio Circles* (ORCs), in the ASKAP EMU pilot survey, that did not correspond to any types of object or artifacts known to have similar morphological features. As it is extremely likely that new discoveries are still waiting to be found in such data deluge, astronomers have started to explore ML-based methods to automatically search for objects with peculiar morphologies. In this process, various methods were proposed, allowing to rediscover previously identified anomalies (including the first detected ORCs) and identify completely new objects (Gupta et al., 2022; Lochner et al., 2023).

In this context, two major methodologies were used. Gupta et al. (2022) and Mostert et al. (2021) employed rotation and flipping invariant self-organizing maps (SOMs) to search for anomalies in the ASKAP EMU pilot and LOFAR LoTSS survey data, respectively. Both analysis used images of fixed size (approximately 1’ to 5’, ~150 pixels per size), centered around previously catalogued radio sources. The Euclidean distance from each “representative” image in the SOM lattice was used as an “anomaly proxy”, e.g. anomalous images have larger Euclidean distances from their closest SOM template image.

Segal et al. (2023) used a coarse-grained complexity metric as an “anomaly” proxy to detect peculiar objects in the ASKAP EMU pilot survey. Their method is based on the idea that image frames containing complex and anomalous objects have a higher Kolmogorov complexity compared to ordinary frames. In contrast to the previously mentioned methods, Segal et al. (2023) conducted a blind search by sliding fixed image frames of size 256×256 pixels (~12 arcmin) through the entire map, rather than focusing on frames centered around known source positions. An approximated complexity estimation for each frame was then computed from the compression file size (using the gzip algorithm) of smoothed and resized frames. This allowed the authors to obtain a catalogue of peculiar sources at different reliability levels, corresponding to different complexity threshold choices. The complexity metric is conceptually simple and fast to compute, which is undoubtedly a positive aspect of this method. However, as noted by Mostert et al. (2021), the complexity metric may not fully capture the morphological features of the sources present in the images.

A potential limitation of “source-centric” approaches could be their reliance on catalogues created with traditional source finding algorithms, which are known to have a higher likelihood of missing diffuse sources (a primary target in anomaly searches). Nevertheless, existing studies successfully manage to identify new anomalous sources in their datasets. Mostert et al. (2021) also noted that their method is not fully sensitive to detect anomalies at angular scales much smaller than the chosen image size (100 arcsec in their work). The choice of the frame size is an aspect that certainly affects “blind” anomaly searches as well.

In this work, we aim to carry out a blind anomaly search study using a different method, which relies on image features extracted by trained self-supervised models. Details on the dataset used and the methodology are provided in the following paragraphs.

### 5.1 Dataset

For this analysis, we considered the `hulk_emupilot` dataset (55,774 images) described in Section 2.2. We annotated through visual inspection approximately 10% of the data (5800 images) using the following set of labels:

- `BACKGROUND`: If the image is purely background noise, e.g. no sources are visible. Typically, this label is set for frames located at the map borders;
- `COMPACT`: if point sources or compact sources comparable with the synthesized beam size (say <10 times the beam) are present. Double or triple sources with point-like components also fall into this category;
- `EXTENDED`: if any extended source is visible, e.g. a compact source with extension >10 × beam;
- `RADIO-GALAXY`: if any extended source is visible with a single- or multi-island morphology, suggesting that of a radio galaxy (e.g. core + lobes);
- `DIFFUSE`: if any diffuse source is visible, typically having small-scale (e.g. <few arcmin) and roundish morphology;
- `DIFFUSE-LARGE`: if any large-scale (e.g. covering half of the image) diffuse object with irregular shape is visible;
- `FILAMENT`: if any extended filamentary structures is visible;

https://doi.org/10.1017/pasa.2024.84 Published online by Cambridge University Press

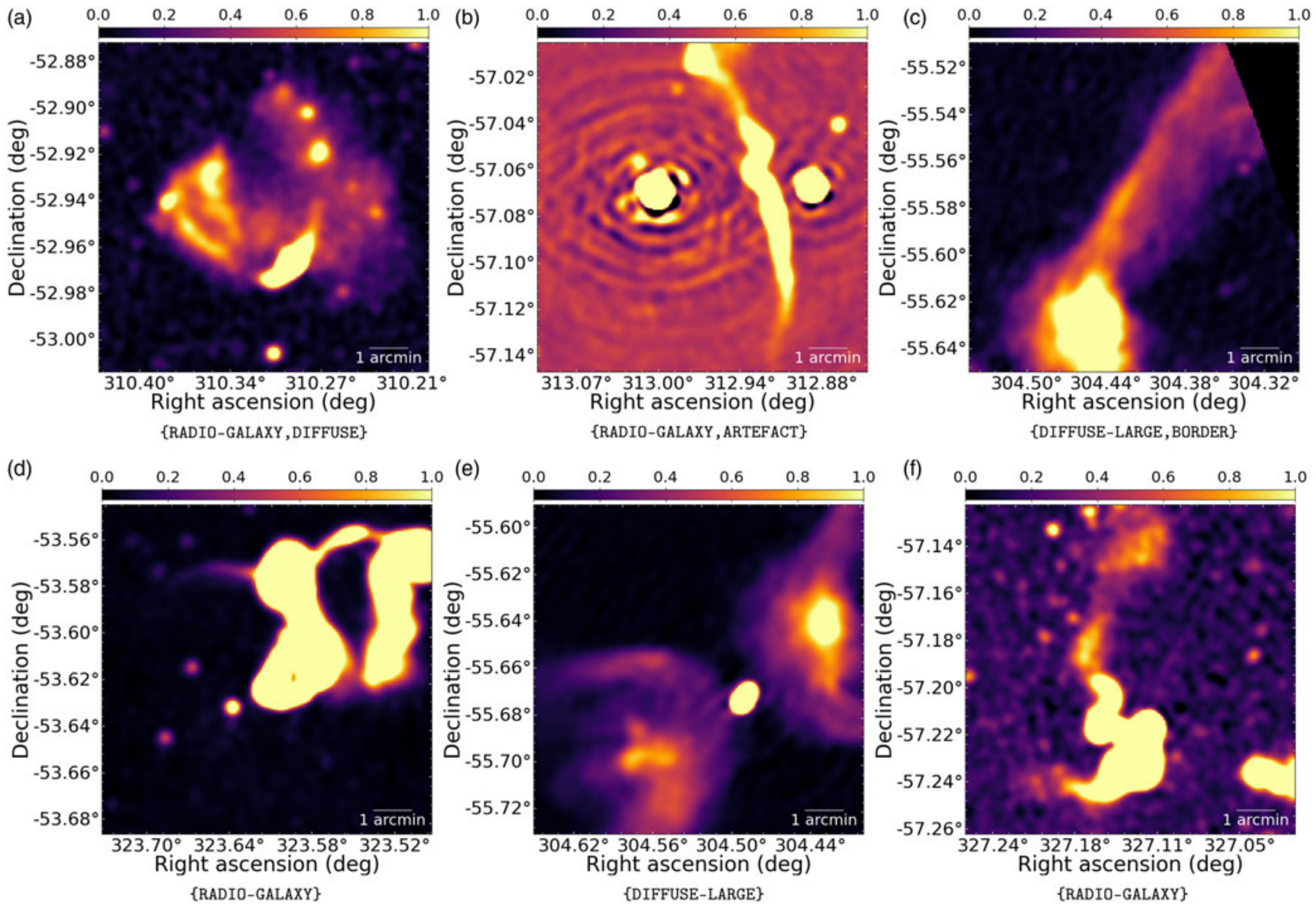


**Figure 11.** Sample images from the `hulk_emupilot` dataset, labelled as PECULIAR and COMPACT. The other assigned labels are reported below each frame. A zscale transform was applied to all images for visualization scopes.

- ARTEFACT: if any ring-shaped or ray-like artefact is visible, e.g. typically around bright resolved sources;
- PECULIAR: if any object is found with peculiar/anomalous morphology;
- MOSAICKING: if any residual pattern of the mosaicking process used to produce the image is present.

More than one label can be assigned to each image, depending on the object/features the user recognize in the image.

A total of 428 peculiar frames were selected through visual inspection starting from a list of 1198 peculiar frames identified in Segal et al. (2023) with a complexity metric analysis and from a catalogue of 361 peculiar sources reported in Gupta et al. (2024). In Figure 11 we show examples of peculiar images from the dataset with their annotation labels.

### 5.2 Anomaly analysis

The data representation variables are each sensitive to different features of the images, including details (e.g. the presence of image borders or artifacts, background noise or mosaicking patterns, compact source density, etc) that are not relevant for the anomaly search task. We tried to limit the dependency on background features with the *RandomThresholding* augmentation, but the model was not fully made invariant with respect to the other aspects. For this reason, we carried out a feature selection analysis, aiming to explore and select features that are mostly correlated with the presence of objects with diffuse or extended morphology. We divided the labelled set of images into two groups: “interesting” frames include images labelled as {EXTENDED,DIFFUSE,DIFFUSE-LARGE}, while “ordinary” frames include the rest of labelled images, mostly hosting only compact sources or artifacts around them. We then trained a LightGBM[k] (Ke et al., 2017) classifier to classify the two groups with all representation features (512) as inputs. A subset of available data was reserved as a cross-validation set for model training early stop. Using shallow decision trees (`max_depth=2`) and default LightGBM parameters (`num_leaves=32`, `min_data_in_leaf=20`), we obtained a classification F1-score of 75.3%. In Figure A.2 we report a plot showing the feature importance returned by the LightGBM trained model. As one can see, a small set of features are identified as the most powerful for selecting interesting frames. We therefore

[k] LightGBM is a high-performance gradient boosting framework based on decision tree algorithm, particularly suited for classification tasks on tabular data. More details are available at https://lightgbm.readthedocs.io/en/latest/index.html.

https://doi.org/10.1017/pasa.2024.84 Published online by Cambridge University Press

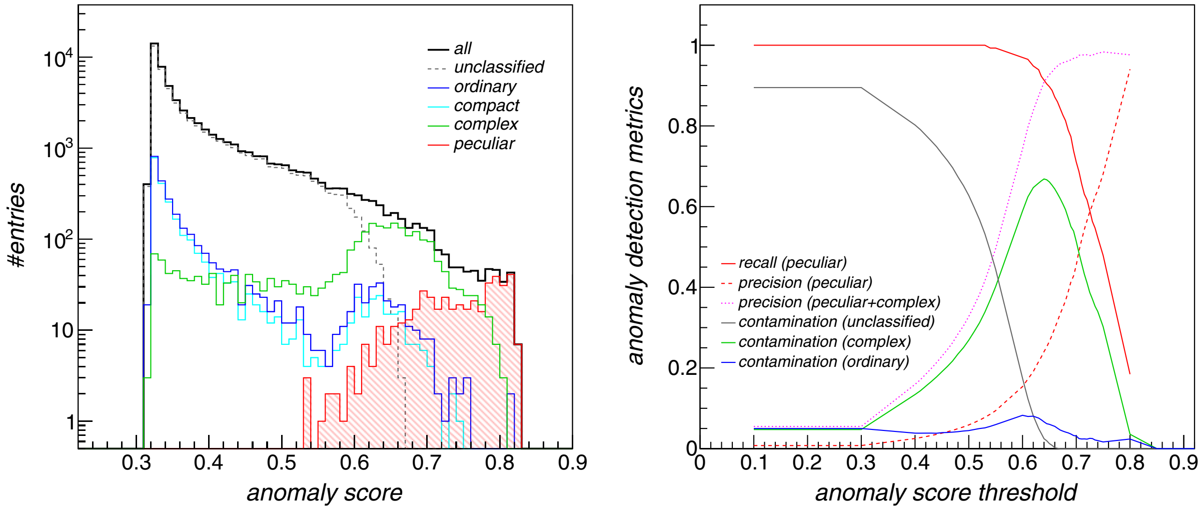


**Figure 12.** Left: Anomaly score of frames contained in the `hulk_emupilot` dataset, shown as black solid histogram, found with the *Isolation Forest* algorithm over top-10 feature data. Unclassified frames are shown with a dashed line. Red filled histogram are the scores of peculiar frames. Ordinary frames (e.g. hosting only compact or artefacts) are shown in blue, pure compact frames in light blue, while frames not tagged as peculiar that host complex sources or structures (`EXTENDED`, `DIFFUSE`, `DIFFUSE-LARGE`, `RADIO-GALAXY`) are shown in green. Right: Anomaly detection metrics (recall, precision, contamination) as a function of the applied anomaly score threshold. Red solid and dashed lines indicate the recall and precision achieved on peculiar frame detection. Purple dotted line is the precision obtained over both peculiar and complex frames. The other solid coloured lines indicate the fraction of unclassified (black line), complex (green line) and ordinary frames contaminating the selected anomaly sample.

carried out the following data exploration and unsupervised analysis, restricting the parameter set to the top-15 ranked variables in importance.

In Figure 3(a) we report a two-dimensional projection of the top-15 variables produced with the UMAP algorithm as a function of the image noise rms level in logarithmic scale (coloured z-axis). As can be seen, the obtained representation shows a residual dependency on physical image parameters, such as the noise rms, that cannot be fully removed by the augmentation scheme currently adopted. In the other panels of Figure A.3 we report the same projection for unlabelled (gray markers) and labelled data, shown with coloured markers. Interestingly, frames that were labelled as peculiar or complex (e.g. containing extended/diffuse objects or artifacts) tend to cluster in specific areas of the projected feature space, also related with higher noise areas, while ordinary frames are uniformly spread in the feature space. Other higher noise areas present in Figure 3(a) seem related to frames that are closer to the mosaic edges or having artifacts (see Figure 3(b)).

We searched for peculiar frames using the *Isolation Forest* (IF) (Liu et al., 2008) outlier detection algorithm[1]. We tuned these IF hyperparameters using the annotated dataset:

- `contamination`: The proportion of outliers in the data set. We scanned these values: ‘auto’, 0.001, 0.01, 0.1.
- `max_samples`: The number of samples to draw from the training data to train each base estimator. We scanned these values: ‘auto’, 0.001, 0.01, 0.02, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0.

Scans were repeated for different choices of importance ranked feature sets: `top2`, `top5`, `top10`, and `top15`. A number of 200 base estimators were used in the tree ensemble. Other IF parameters were set to defaults. Best classification results were obtained with a smaller fraction of samples (`max_samples`=0.02) and `contamination`=0.001.

We then ran the IF algorithm in an unsupervised way with tuned parameters and obtained an anomaly score for each dataset frame. The anomaly score ranges from 0 to 1, with most anomalous data expected to have values close to 1. In Figure 12 (left panel) we report the distribution of IF anomaly scores of all frames contained in the `hulk_emupilot` dataset, shown as a black solid histogram, found over top-10 feature data. Unclassified frames are shown with a dashed line. The red filled histogram indicates the labelled peculiar frames. Ordinary frames (e.g. hosting only compact or artifacts) are shown in blue, pure compact frames in light blue, while complex frames (e.g. hosting extended or diffuse structures, not labelled as peculiar) are shown in green.

Following Section 2.5, we computed the anomaly detection metrics (peculiar frame recall and precision, non-peculiar frame contamination) as a function of the applied anomaly score threshold. Peculiar frame recall and precision are reported in Figure 12 (right panel) as a function of the applied anomaly score threshold for top-10 feature data, respectively shown with solid and dashed red lines. We also computed the precision in classifying detected frames as either peculiar or complex, shown with a dotted purple line. The other solid coloured lines indicate the fraction of unclassified (black line), complex (green line) and ordinary frames contaminating the selected anomaly sample. In Table 4 we summarized the metrics obtained for different feature sets for the anomaly score threshold that provides the best peculiar recall/precision compromise (e.g. the score at which recall and precision curves cross in Figure 12(b)). Best detection performances (∼60%) are obtained with the top-5 features, but the top-10 feature set currently provides the smallest contamination fraction of ordinary

[1] *Isolation Forest* is an unsupervised decision-tree-based algorithm for outlier detection in tabular data, that works by randomly selecting a feature and a random split value to isolate data points in a binary tree. It identifies outliers as instances that require fewer splits to be isolated, exploiting the inherent rarity of anomalies in a dataset.

https://doi.org/10.1017/pasa.2024.84 Published online by Cambridge University Press

**Table 4.** Peculiar frame detection metrics obtained with the *Isolation Forest* algorithm over selected feature sets (column (1)) when using an anomaly score threshold (reported in column (2)) that provides the best compromise in terms of peculiar frame recall and precision, respectively shown in columns (3) and (4). The precision relative to joint peculiar and complex frames is shown in column (5). The fractions of complex and ordinary frames contaminating the predicted anomalous sample are shown in columns (6) and (7).

| | | $\mathcal{R}$ | $\mathcal{P}$ | $\mathcal{P}_{pec+complex}$ (%) | $\mathcal{C}_{complex}$ | $\mathcal{C}_{ordinary}$ |
|---|---|---|---|---|---|---|
| Features | Thr. | (%) | (%) | (%) | (%) | (%) |
| `top2` | 0.700 | 55.6 | 59.8 | 93.5 | 33.7 | 6.5 |
| `top5` | 0.750 | 61.2 | 63.0 | 93.0 | 30.0 | 7.0 |
| `top10` | 0.725 | 59.1 | 58.7 | 97.4 | 38.7 | 2.6 |
| `top15` | 0.660 | 57.7 | 58.7 | 95.5 | 36.8 | 4.5 |

frames ($<$3%). When considering both peculiar and complex frames, the precision increases to 97%.

### 5.3 *Astronomer-in-the-loop*

It is worth to note that the source peculiarity concept is rather subjective and may depend on the scientific domain of interest. For instance, a fraction of complex frames may well be considered as truly peculiar in specific analysis, and, on the other hand, missed peculiar frames may be considered not as relevant in other contexts. For this reason, an additional "human-in-the-loop" processing stage has to be applied to our list of candidate anomalies to create a refined sample that better fits scientific needs.

For the sake of demonstration, we integrated our dataset in the `astronomaly` package[m] (Lochner & Bassett, 2021). This allowed us to run an active learning process from a web interface in which users can personalize and sort the list of anomalous frames on the basis of the computed score and also their expressed preferences, such as how peculiar a frame is judged on a scale of 1 to 5. A screenshot of the `astronomaly` UI for our dataset is shown in Figure A.4.

We plan to integrate in the future the full pipeline (feature extraction, anomaly detection, active learning loop) as a supported application within the *caesar-rest* service[n] (Riggi 2021), and extend the web UI with missing functionalities (e.g. image filtering/exporting, model importing, configuration options, etc). In this study, we limited ourselves to primarily quantify the ordinary frame rejection power that can be currently achieved with self-supervised features, as this will largely impact the time needed to visually inspect the anomaly candidates in human-in-the-loop approaches to form the final anomaly sample.

## 6. Summary

In this study, we investigated the potential of self-supervised learning for analysing radio continuum image data produced by SKA precursors. Specifically, we have used the SimCLR contrastive learning framework to train deep network models on large sets of unlabelled images extracted from the ASKAP EMU pilot and SARAO MeerKAT GPS surveys, either randomly selected or centred around catalogued extended source positions. The trained encoder network, based on the *ResNet18* architecture, was used as a feature extractor and fine-tuned for three distinct downstream tasks (source detection, morphology classification, and anomaly detection) over test datasets comprising thousands of annotated images from other radio surveys (VLA FIRST, ASKAP Early Science, ATCA Scorpio surveys). Notably, some of these test datasets were purposefully created for this work.

All trained models, including both the source code and network weights, have been publicly released. These represent a first outcome of this work, as they can be viewed as prototypal radio foundational models, available to be used in future applications for multiple scopes:

- to extract feature parameters from new radio survey images and perform data inspection, unsupervised classification or outlier detection analysis (as demonstrated in Section 5);
- to serve as pre-trained backbone components of more complex models designed for source classification, detection or other tasks (e.g. source property characterization), eventually refined over new labelled datasets (as demonstrated in Sections 3 and 4).

The analyses we performed in this work attempted to address various open questions in this field, paving the way for future analyses:

- Do we observe any advantages stemming from self-supervised models trained on easily constructed "random" survey datasets compared to costly-to-compile "source-centric" datasets?
- How does self-supervised learning on radio data compare in performance to models pre-trained on extensive non-radio datasets, such as *ImageNet*?
- Is it feasible to enhance existing radio source detectors utilizing deep networks through radio self-supervised pre-training?

We found that using uncurated large collections of unlabelled radio images randomly extracted from SKA precursor surveys resulted in significantly improved performances (approximately 5%) in both radio source detection and classification tasks, compared to curated (albeit smaller) image samples extracted around extended source catalogues. This indication, primarily attributed to the augmented number of accessible images achievable with uncurated collections, is highly encouraging, as it suggests that certain aspects of source analysis can be enhanced even without investing numerous work months in catalogue production.

The advantages gained from self-supervised pre-training on radio data, compared to non-radio data, are notably significant (exceeding 10%) in both source classification and detection tasks. However, when contrasting our findings with fully supervised models trained from scratch, we observed that these benefits are only relevant with small labelled datasets (on the order of a few hundred images). This is certainly a positive aspect, considering that many available annotated datasets (such as *MiraBest* or similar radio galaxy classification datasets) typically fall within this size range. Nevertheless, in order to observe a substantial impact on larger datasets, it becomes imperative to improve both the self-supervised pre-training dataset and the model itself.

[m] https://github.com/MichelleLochner/astronomaly
[n] https://github.com/SKA-INAF/caesar-rest

We have identified some areas of developments to be made in the near future to improve source analysis performance, and overcome the limitations encountered in this study. Firstly, we plan to increase the size of our pre-training `hulk` datasets, by leveraging the massive amount of unlabelled image data being delivered by large area surveys, such as ASKAP EMU, the Very Large Array Sky Survey (VLASS) (Lacy et al., 2020), or the LOFAR Two-metre Sky Survey (LoTSS) (Shimwell et al. 2017) surveys. In this context, to reduce the computational load during training, it is crucial to explore effective and automated strategies for constructing semi-curated large-scale pre-training datasets, potentially comprising millions of images. This step may require the development of specialized algorithms to filter or weight image frames included in the pre-training dataset, aiming to maximize the balance between ordinary and complex objects "seen" by the model.

Additionally, we have already started to train larger architectures and recent state-of-the-art self-supervised frameworks, particularly those based on Vision Transformers (ViTs), over the same datasets produced for this study. Results will be compared against the SimCLR baseline and presented in a forthcoming paper.

**Acknowledgement.** This scientific work uses data obtained from Inyarrimanha Ilgari Bundara/the Murchison Radio-astronomy Observatory. We acknowledge the Wajarri Yamaji People as the Traditional Owners and native title holders of the Observatory site. CSIRO's ASKAP radio telescope is part of the Australia Telescope National Facility (https://ror.org/05qajvd42). Operation of ASKAP is funded by the Australian Government with support from the National Collaborative Research Infrastructure Strategy. ASKAP uses the resources of the Pawsey Supercomputing Research Centre. Establishment of ASKAP, Inyarrimanha Ilgari Bundara, the CSIRO Murchison Radio-astronomy Observatory and the Pawsey Supercomputing Research Centre are initiatives of the Australian Government, with support from the Government of Western Australia and the Science and Industry Endowment Fund.

This work made use of PLEIADI, a computing infrastructure installed and managed by INAF.

**Funding statement.** This work received funding from the INAF CIRASA and SCIARADA projects.

**Competing interests.** None.

**Data availability statement.** The software code used in this work is publicly available under the GNU General Public License v3.0[o] on the GitHub repository https://github.com/SKA-INAF/sclassifier/. The trained model weights have been made available on Zenodo repository at https://doi.org/10.5281/zenodo.12636593.

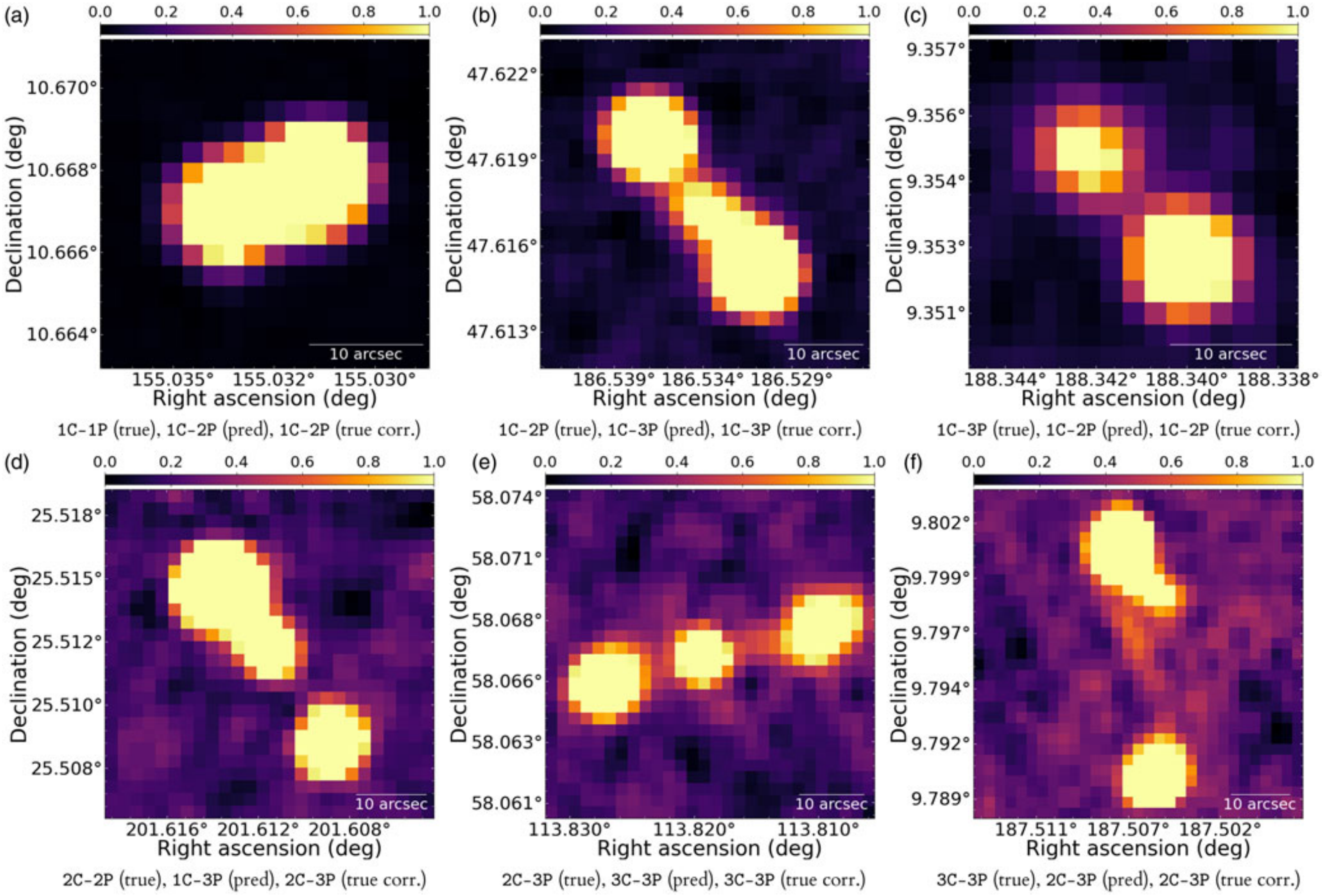
## References

Aniyan, A. K., & Thorat, K. 2017, ApJS, 230, 20
Banfield, J. K., et al. 2015, MNRAS, 453, 2326
Becker, R. H., et al. 1995, MNRASx, 450, 559
Bengio, Y., et al. 2013, IEEE Trans. PATTERN Anal. Mach. Intell., 35, 1798
Bonaldi, A., et al. 2021, MNRAS, 500, 3821
Bordiu, C., et al. 2024, A&A, submitted
Bordiu, C., et al. 2023, (eds) Machine Learning for Astrophysics. ML4Astro 2022. Astrophysics and Space Science Proceedings, vol 60. Springer, Cham. https://doi.org/10.1007/978-3-031-34167-0_13
Boyce, M. M., et al. 2023, PASA, 40, e027
Chen, T., et al. 2024 Proceedings of the 37th International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020
Cornu, D., et al. 2024, A&A, 690, A211
Deng, J., et al. 2009, Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pp. 248–255, https://doi.org/10.1109/CVPR.2009.5206848
Dewdney, P., et al. 2016, SKA1 SYSTEM BASELINE DESIGN V2, SKA-TEL-SKO-0000002
Galvin, T. J., et al. 2020, MNRAS, 497, 2730
Goedhart, S., et al. 2024, MNRAS, 531, 649
Grill, J.-B., et al. 2020, In Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20). Curran Associates Inc., Red Hook, NY, USA, Article 1786, 21271–21284.
Gupta, N., et al. 2024, PASA, 41, e027
Gupta, N., et al. 2024, PASA, 41, e001
Gupta, N., et al. 2023, PASA, 40, e044
Gupta, N., et al. 2022, PASA, 39, e051
He, K., et al. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.
Hossain, M. S., et al. 2023, Procedia Computer Science 222, 601
Hotan, A., et al. 2021, PASA, 38, E009
He, K., et al. 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 2017, pp. 2980–2988, doi: 10.1109/ICCV.2017.322.
Johnston, S., et al. 2008, ExA, 22, 151
Ke, G., et al., 2017, Adv. Neural Inform. Process. Syst., 30, 3146.
Koribalski, B. S., et al. 2021, MNRAS, 505, L11
Lacy, M., et al. 2020, PASP, 132, 035001
Lao, B., et al. 2023, A&C, 44, 100728
Lochner, M., Bassett, B. A. 2021, A&C, 36, 100481
Lochner, M., et al. 2023, MNRAS, 520, 1439
Liu, F. T., et al. 2008, 8th IEEE International Conference on Data Mining, page 413–422, https://doi.org/10.1109/ICDM.2008.17
Liu, X., et al. 2023, IEEE Trans. Knowl. Data Eng., 35, 857
Lukic, V., et al. 2018, MNRAS, 476, 246
Ma, Z., et al. 2019, ApJS, 240, 34
McInnes et al. 2018, J. Open Source Softw., 3(29), 861
Mohale, K., & Lochner, M. 2024, MNRAS, 530, 1274
Mostert, R. I. J., et al. 2021, A&A, 645, A89
Mostert, R. I. J., et al. 2022, A&A, 668, A28
Ndung'u, S., et al. 2023, NewAR 97, 101685
Norris, R. P., et al. 2011, PASA, 28, 215
Norris, R. P., et al. 2021a, PASA, 38, e046
Norris, R. P., et al. 2021b, PASA, 38, e003
Pennock, C. M., et al. 2022, MNRAS, 515, 6046
Porter, F. A.M., & Scaife, A. M.M. 2023, RAS Techniq. Instrum., 2, 293
Ralph, N. O., et al. 2019, PASP, 131, 108011
Riggi, S., et al. 2021, A&C, 37, 100506
Riggi, S., et al. 2023, A&C, 42, 100682
Segal, G., et al. 2023, MNRAS, 521, 1429
Slijepcevic, V. I., et al. 2022, MNRAS, 514, 2599
Slijepcevic, V. I., et al. 2024, RAS Techniq. Instrum., 3, 19
Shimwell, T. W., et al. 2017, A&A, 598, A104
Tang, H., et al. 2019, MNRAS, 488, 3358
Tsung-Yi, L., et al. 2014, Microsoft COCO: Common Objects in Context. CoRR, abs/1405.0312, http://arxiv.org/abs/1405.0312
Umana, G., et al. 2015, MNRAS, 454, 902
Umana, G., et al. 2021, MNRAS, 506, 2232
Yu, Y., et al. 2022, MNRAS, 511, 4305
Wright, E. L., et al. 2010, AJ, 140, 1868
Wu, C., et al. 2019, MNRAS, 482, 1211
Zhang, Z., et al. 2022, PASP, 134, 064503

[o] https://www.gnu.org/licenses/gpl-3.0.html

# Appendix

## A. Supplementary plots



(a) 1C-1P (true), 1C-2P (pred), 1C-2P (true corr.)

(b) 1C-2P (true), 1C-3P (pred), 1C-3P (true corr.)

(c) 1C-3P (true), 1C-2P (pred), 1C-2P (true corr.)

(d) 2C-2P (true), 1C-3P (pred), 2C-3P (true corr.)

(e) 2C-3P (true), 3C-3P (pred), 3C-3P (true corr.)

(f) 3C-3P (true), 2C-3P (pred), 2C-3P (true corr.)

**Figure A.1.** Examples of sources from the RGZ test dataset that were misclassified by the trained source classifier (`hulk_smgps` pre-trained and frozen backbone model) due to an incorrect true class label provided in the dataset (mislabelling). The true and predicted class labels are reported below each frame. In many cases, the model indeed correctly predicted the expected true classification (denoted as "true corr." below each frame).

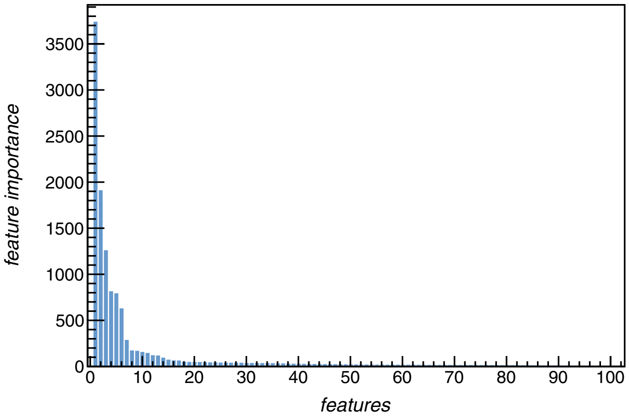
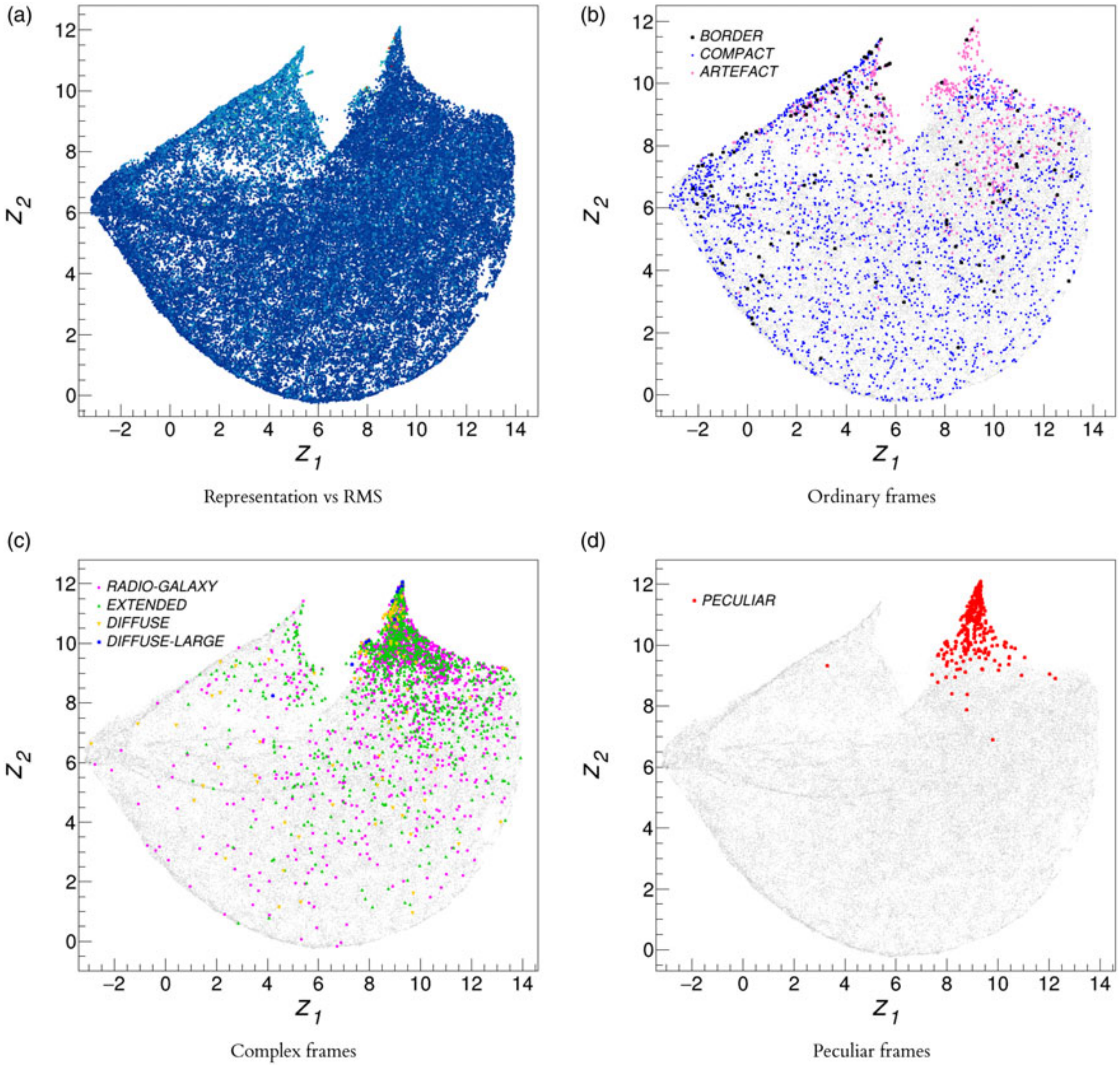


**Figure A.2.** Feature importance obtained with a LightGBM classifier trained on `hulk_emupilot` data representation, for the classification of interesting against ordinary images.

https://doi.org/10.1017/pasa.2024.84 Published online by Cambridge University Press



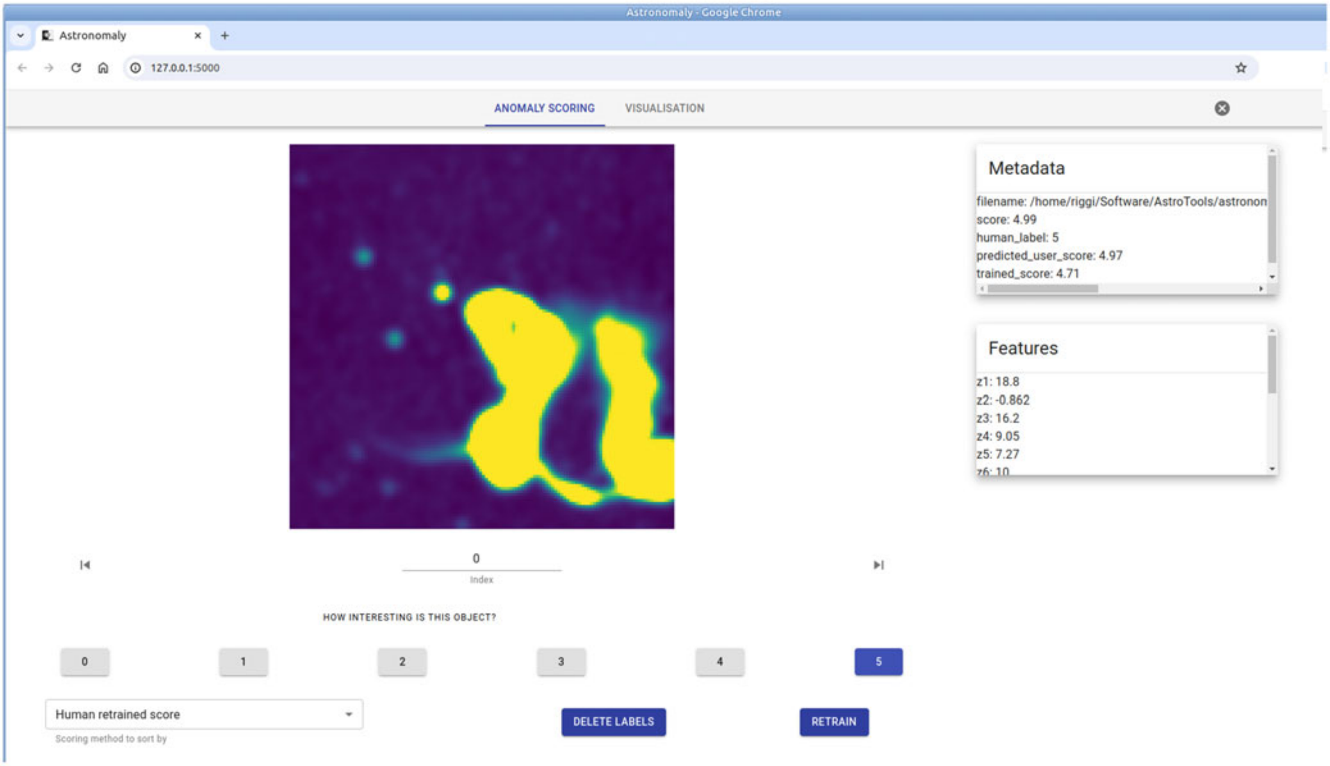
Representation vs RMS

Ordinary frames

Complex frames

Peculiar frames

**Figure A.3.** Figure 14: 2D UMAP projection of the top-15 selected features from the data representation vector produced by the trained SimCLR model on the `hulk_emupilot` dataset as a function of the image noise RMS level in logarithmic scale (z-scale axis). Red markers correspond to image with higher RMS levels, while blue markers to low noise RMS levels. Left: 2D UMAP projection of the top-15 selected features for unclassified frames (gray markers) and labelled frames (coloured markers, as reported in the plot legends). See text for details on label schema.



Metadata

filename: /home/riggi/Software/AstroTools/astronon
score: 4.99
human_label: 5
predicted_user_score: 4.97
trained_score: 4.71

Features

z1: 18.8
z2: -0.862
z3: 16.2
z4: 9.05
z5: 7.27
z6: 10

Human retrained score
Scoring method to sort by

**Figure A.4.** Screenshot of `astronomaly` web UI with list of anomalous frames selected from the `hulk_emupilot` dataset.

https://doi.org/10.1017/pasa.2024.84 Published online by Cambridge University Press