Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
ICSC Centro Nazionale di Ricerca in HPC, Big Data and Quantum Computing

Object detection, self-supervision and XAI using real and synthetic astronomical data

Simone Riggi, Thomas Cecconello, Filomena Bufano

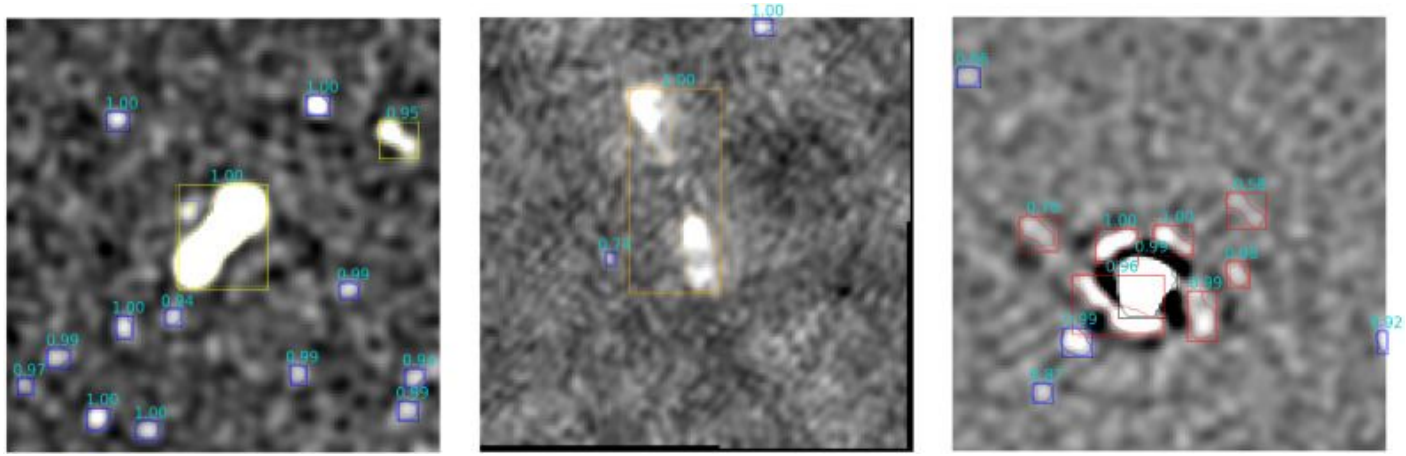
Spoke 3 Technical Workshop, Trieste October 9 / 11, 2023

Use cases

1. Object detector framework for radio source extraction
   a. Simone Riggi, Thomas Cecconello
2. Radio source analysis with supervised and self-supervised learning techniques
   a. Simone Riggi, Thomas Cecconello
3. Unsupervised Clustering of stellar SEDs with XAI
   a. Filomena Bufano

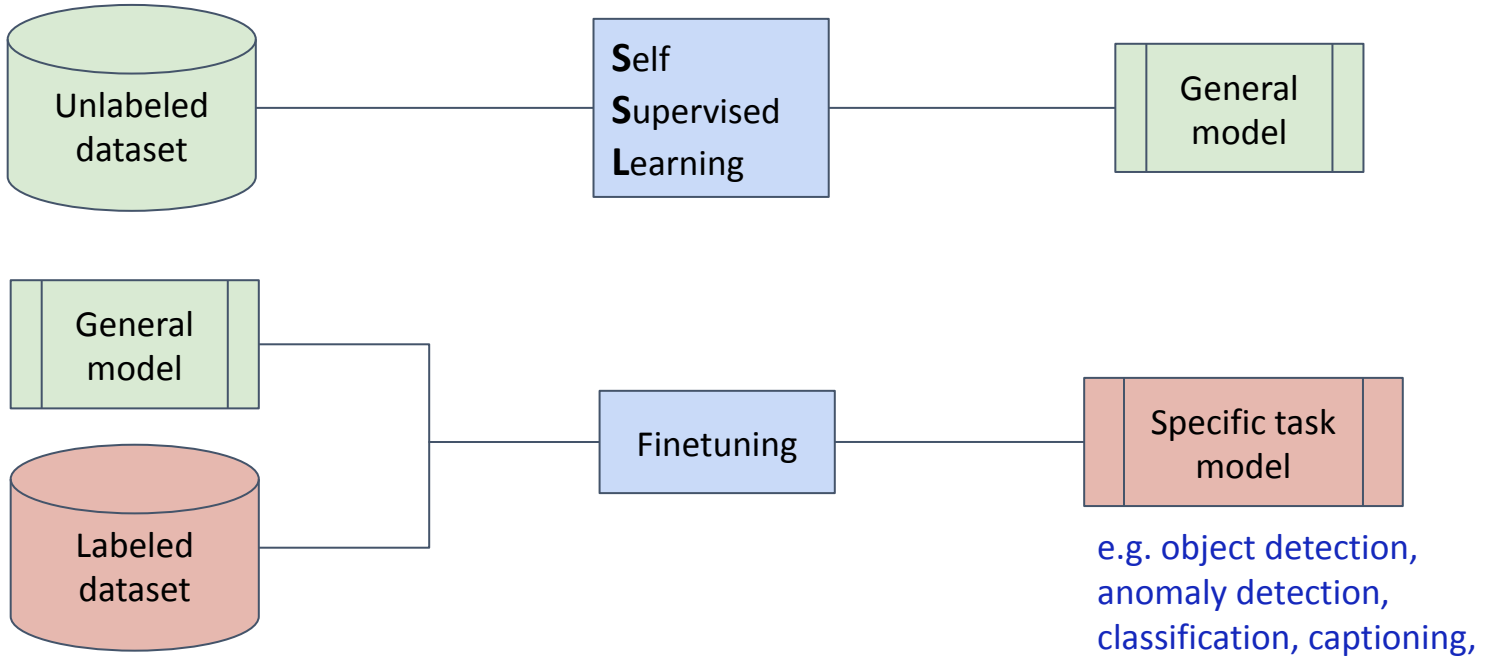
Scientific Rationale - Use case 1

- Develop ML-based source extractor tools able to detect radio sources (multi-island objects, diffuse/extended, artefact, …) missed by traditional finders in SKA and precursor maps

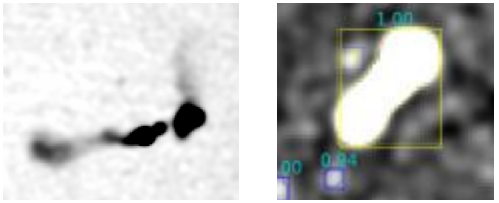


- Reference paper: *"RIGGI, Simone, et al. Astronomical source detection in radio continuum maps with deep neural networks. Astronomy and Computing, 2023, 42: 100682."*
- Survey paper: *"SORTINO, Renato, et al. Radio astronomical images object detection and segmentation: a benchmark on deep learning methods. Experimental Astronomy, 2023, 1-39."*

Scientific Rationale - Use case 2

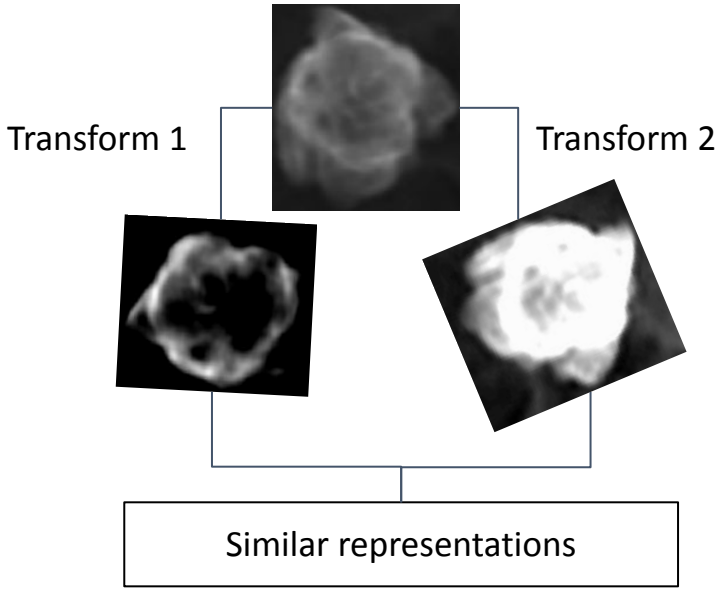


Supervised learning



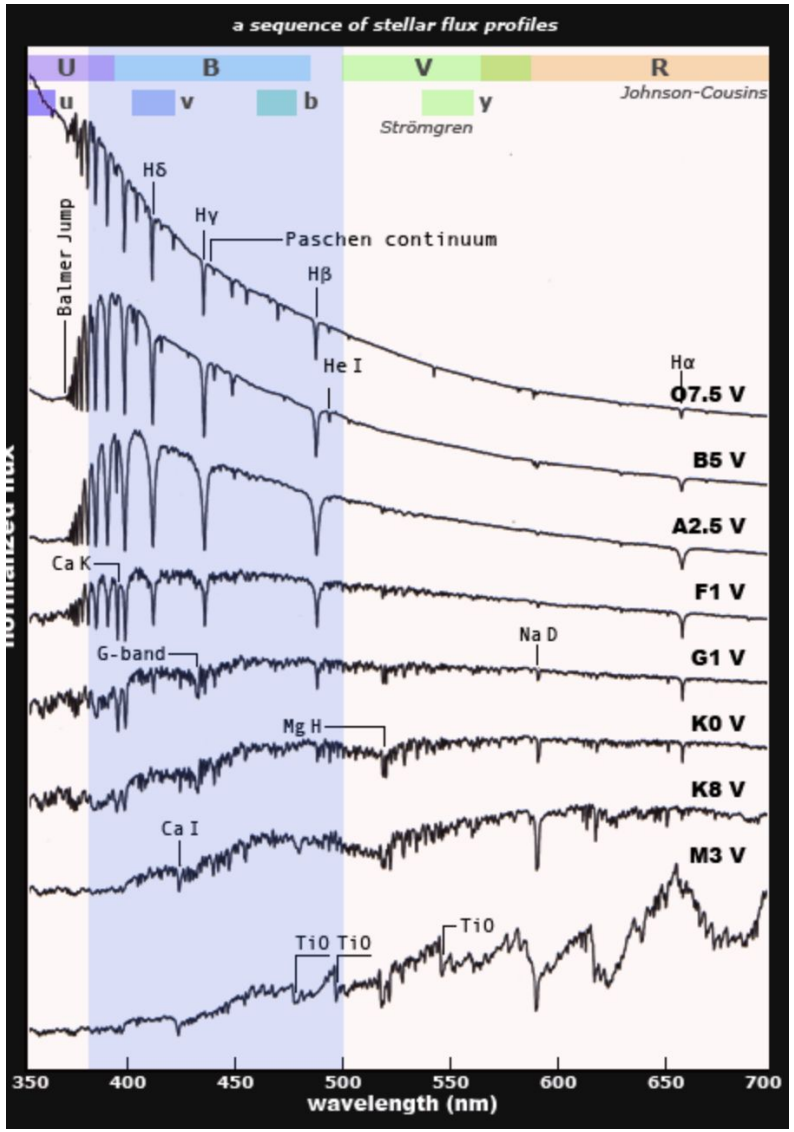
2-islands extended (multi-components)

Self supervised learning



Scientific Rationale - Use case 3

- A surprisingly high percentage of stellar objects in astrophysical catalogues are unclassified
- We aim to offer a method to classify unknown/new sources based on available photometry
- To do that we will train a model on a large collection of synthetic stellar Spectral Energy Distributions (SEDs), finding physically meaningful clusters related to specific physical parameters (e.g. temperature, circumstellar dust)



Technical Objectives, Methodologies and Solutions - Use Case 1

1. Improve model accuracy on specific morphological radio source classes
   a. Increase dataset size and rebalance for specific classes
      i. Add new real data from ASKAP and MeerKAT surveys
      ii. Add synthetic data using diffusion models (link)
   b. Improve image preprocessing (e.g. normalization)
   c. Improve backbone by exploring new methods of pretraining
      e.g. Self Supervised Learning (Use case 2)

2. Multi-node parallelization of prediction stage with MPI and benchmarks
   a. Split large maps in to tiles (in order to fit in GPU memory) and merge border detection

3. Framework update of repository engine (https://github.com/SKA-INAF/caesar-mrcnn)
   a. Tensorflow v1 to v2

4. Application to new SKA and precursor data to create new catalogues/dataset

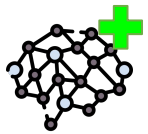
Technical Objectives, Methodologies and Solutions - Use Case 2

1. Benchmark performance using different methodologies:
   a. Transfer learning from model trained on regular images (e.g. ImageNet)
   b. Train a model from scratch with supervised learning on specific tasks
   c. Transfer learning using Self supervised learning on random radio map cutouts
   d. Transfer learning using Self supervised learning on curated dataset (source in the center of the image)

2. Build radio dataset for Self Supervised Learning and test dataset to evaluate performance
   a. Build a dataset with random cutouts from radio tiles
   b. Build unlabeled dataset with sources in the middle of the image

Technical Objectives, Methodologies and Solutions - Use Case 3

1. Generate a training dataset of synthetic stellar SEDs



2. Produce a new model to reduce dimensionality

3. Clusterize the embeddings.

XAI

4. Infer a physical meaning of the embeddings.

Timescale, Milestones

	Oct 23	Apr 24	Oct 24	Apr 25	End
Use case 1	• Code refactored in tensorflow 2 • MPI v1 implementation • SSL model tested	• Large map annotated data ready • Dataset extended • Models availability increased for benchmarking	• Dataset extension completed • Benchmarking complete • MPI implementation completed	• Code&Model release • Results presented in a conference	• Application on new precursor data
Use case 2	• Datasets v1 ready • SSL methods implemented • Beta models ready • Benchmark method defined	• Datasets v2 ready • Benchmark started	• Dataset completed • Benchmark completed	• Code&Model release • Results presented in a conference	
Use case 3	• Data collection started • Clustering and XAI Methods study and selection started	• Dataset v1 of synthetic Stellar SEDs ready	• Dataset v2 ready • Model v1	• Preliminary results shown in a conference	• Trained model and clustering results

Timescale, Milestones and KPIs

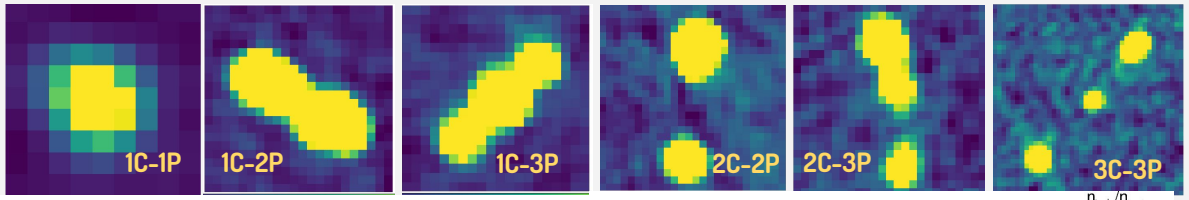
	Timescale	Milestone	KPIs
Use Case 1	Apr 2025	Code & Model release	1
	Apr 2025	Presentation of methods and results in scientific conferences	1
Use Case 2	Apr 2025	Code & Model release	1
	Apr 2025	Presentation of methods and results in scientific conferences	1
Use Case 3	Apr 2025	Presentation of methods and results in scientific conferences	1

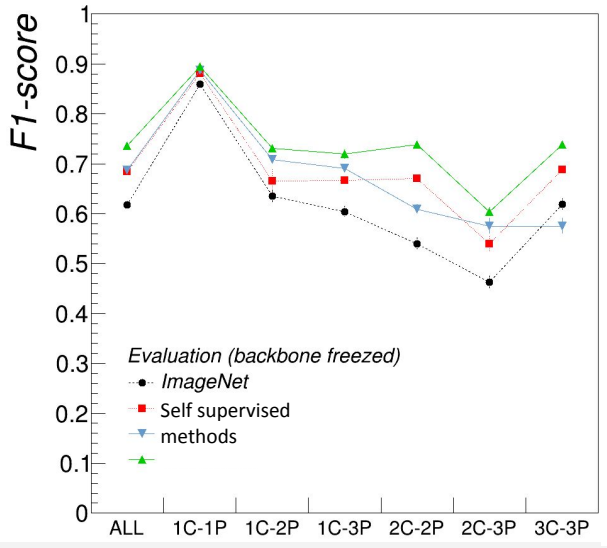
First results - Use cases 2

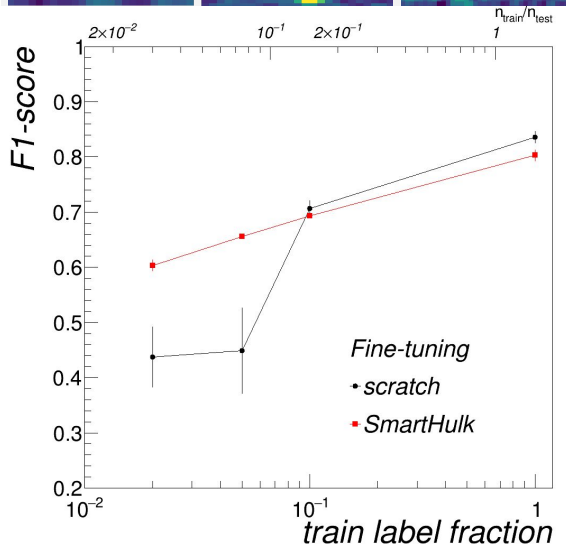
Source Morphology Classification

Radio Galaxy Zoo (RGZ) dataset: #82084

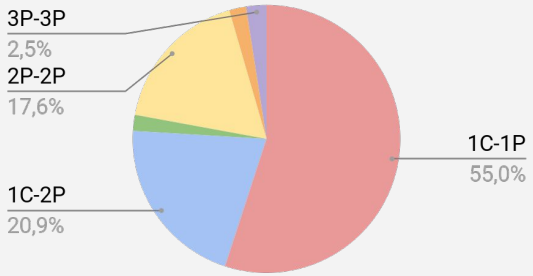
- cutouts centred around sources
- 6 classes (1C-1P, 1C-2P, 1C-3P, 2C-2P, 2C-3P, 3C-3P)
- Balanced datasets for training: #9591 images (train/test: ~1000/600 per class)







Different survey data (FIRST VLA) used for representation testing



- ✓ Better representation wrt ImageNet (+7% ÷ +12%)
- ✓ Random & centred datasets combination works best!
- ✓ Fine-tuning self-supervised model provides better metrics & train times for small labelled datasets

11