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

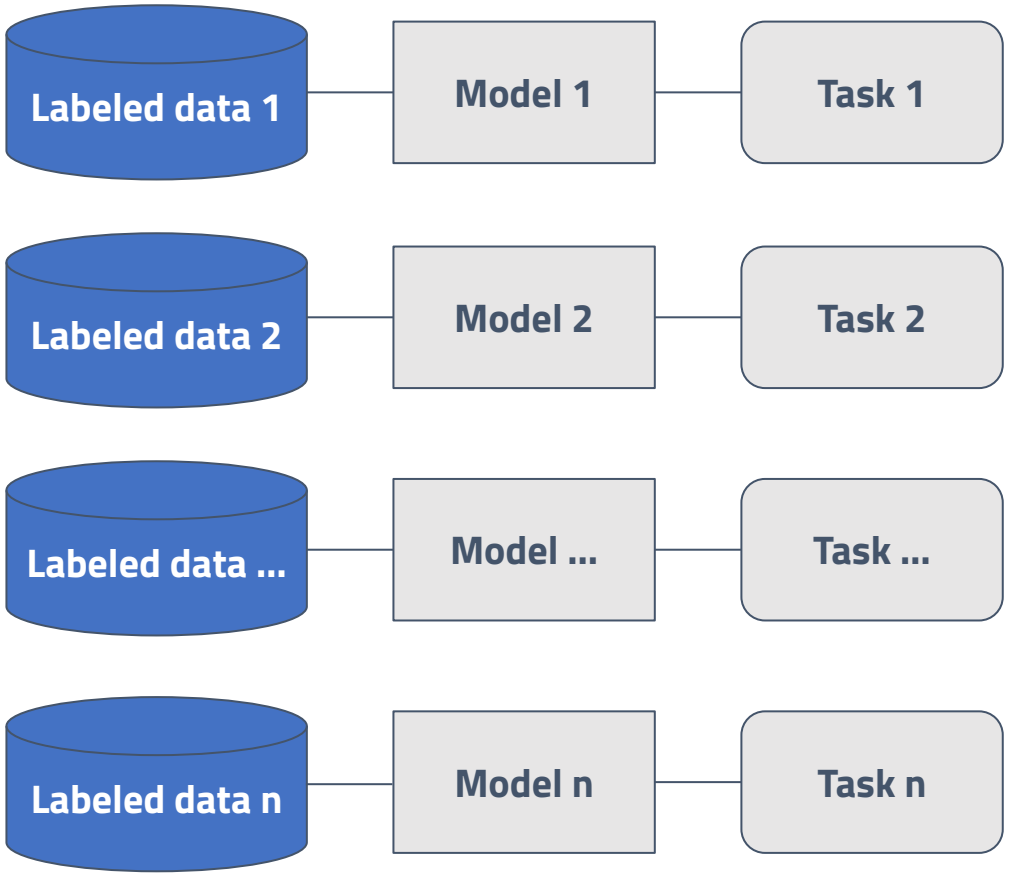
Self-supervision on radio data for source analysis

Thomas Cecconello, Simone Riggi

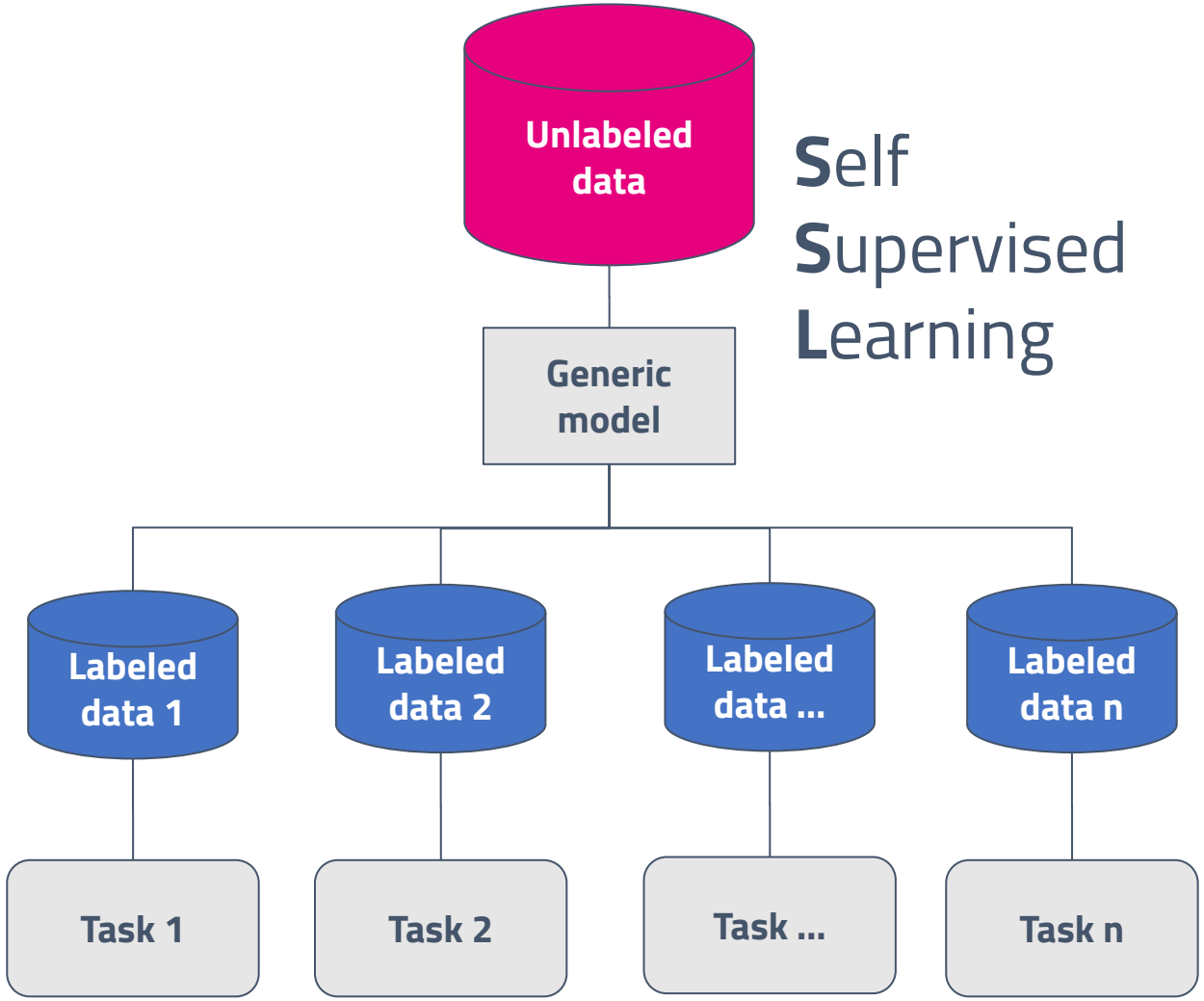
Spoke 3 General Meeting, Elba 5-9 / 05, 2024

Scientific Rationale

Supervised Learning

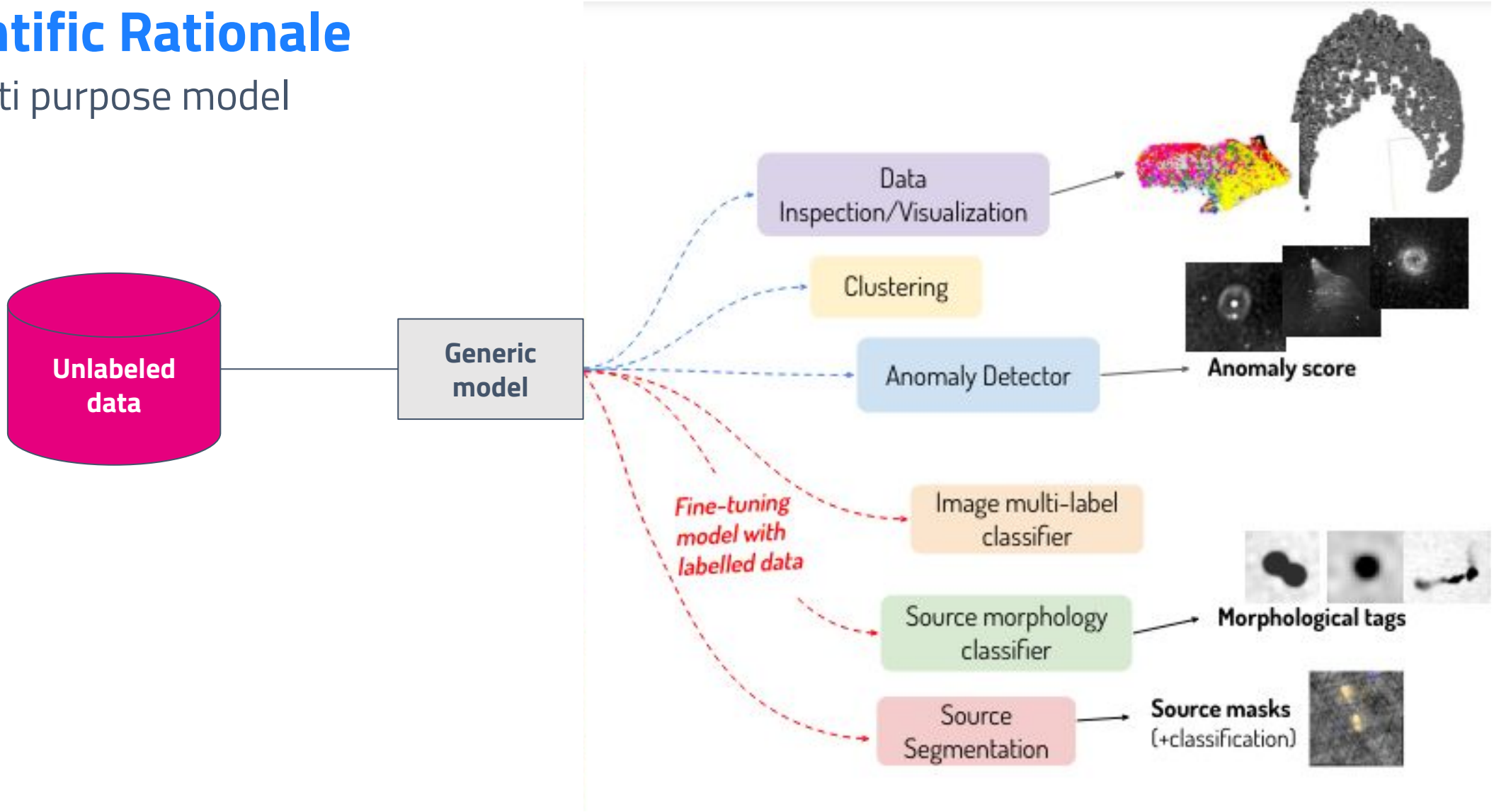


Self Supervised Learning



Scientific Rationale

Multi purpose model



Technical Objectives, Methodologies and Solutions

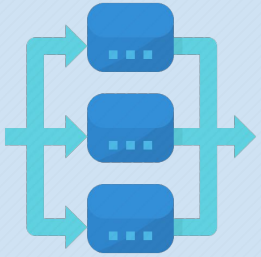
Perform **benchmark** of SSL methods on radio astronomical data

Retrieve and create **datasets**

Develop benchmark **code**

Mask-rcnn code improvement

Parallelize using **MPI** and scale using multiple GPUs



Timescale, Milestones and KPIs

SSL usecase

Datasets v1:
- Huge unlabeled datasets

Code v1:
- BYOL SSL method

Mask-rcnn

Code v2:
- caesar-mrcnn model train (+pre-train)

SSL usecase

Datasets v3:
- More evaluation datasets

Code v3:
- Various code improvements
- Adapted repo to benchmark on LEONARDO

M6 | **M7** | **M8**

Mask-rcnn

Code v1:
- caesar-mrcnn tf2 version

SSL usecase

Datasets v2:
- Evaluation datasets

Code v2:
- Benchmark repository

Mask-rcnn

Code v3:
- caesar-mrcnn parallel version test

Accomplished Work, Results

Run-time on LEONARDO (single A100 64 GB GPU)

- **obtained through ISCRA C**

Pre-train (48 models)

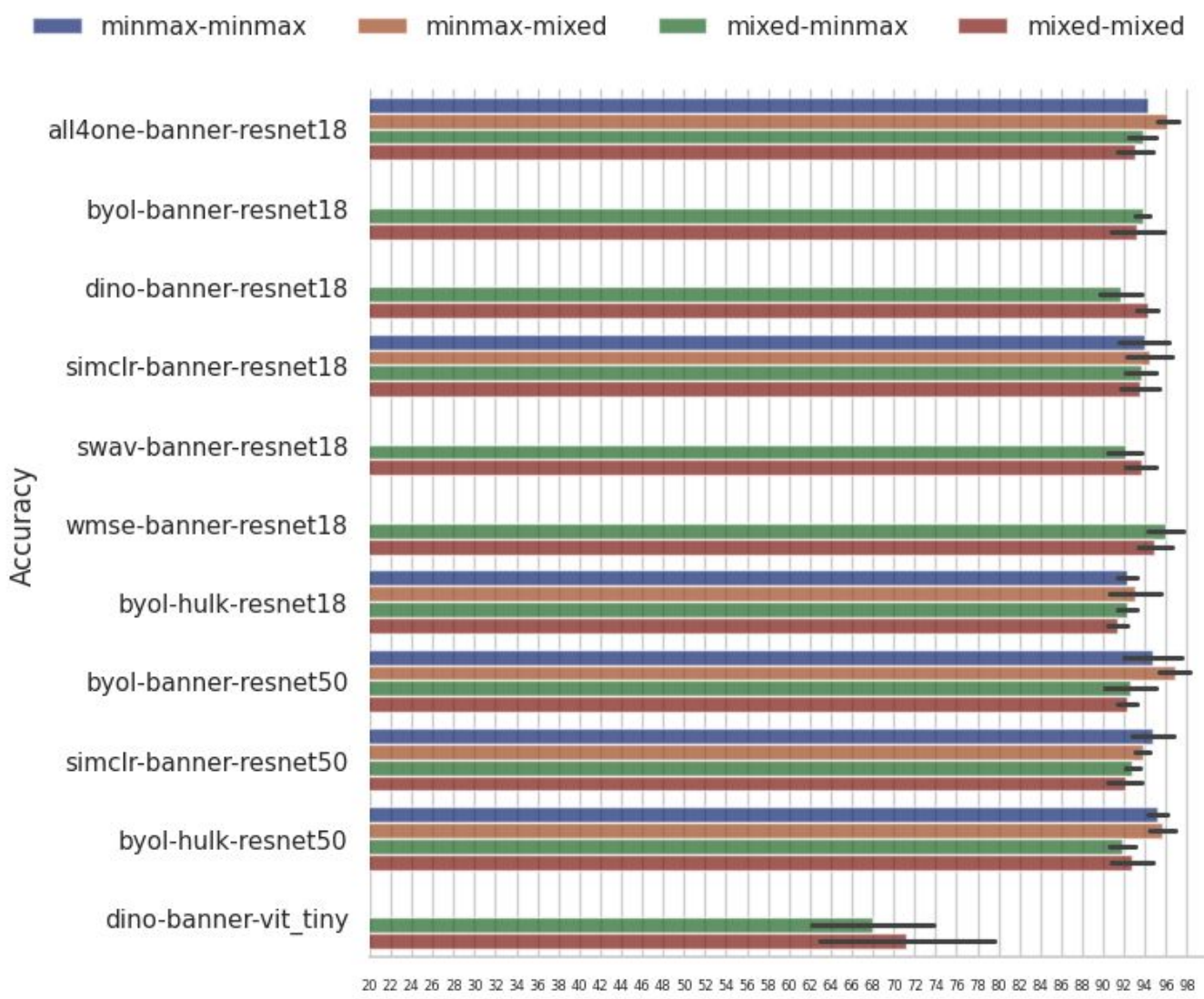
- Resnet18: **~5 Hours**
- Resnet50: **~20 Hours**

Linear evaluation & fine-tuning (40 runs per pre-trained model)

- ~30 min to ~2.5 Hours (depending on test dataset size)

Accomplished Work, Results

Produced first plots to show benchmark results of pre-trained model accuracies on various tasks & datasets



Accomplished Work, Results

- *Repositories*
  - https://github.com/dr4thmos/solo-learn-radio/commits/master/
  - https://github.com/SKA-INAF/caesar-mrcnn-tf2
- *Datasets*
  - https://docs.google.com/spreadsheets/d/1tekXnxrBA3scV7hSIbjm1j-qGSEuwQsc36I-XUpMibs/edit?usp=sharing
  - **2 pretraining dataset, 5 evaluation datasets**
- *Executed runs*
  - **1129 runs, ~12k hours GPU used**
- *Dissemination*
  - **CERAML: INAF-UniMalta workshop on AI for students**
- *Paper*
  - **Self-Supervised Learning benchmark paper in completion**

Next Steps and Expected Results

Short-term (next milestones)

- Continue benchmark and improvement
- Release pretrained models (~30)
- Submit paper to journal or conference

On a longer term

- Strategies for improved and enlarged pre-training datasets
- Training vision transformers on multiple GPUs

Thank you for your attention

BACKUP SLIDES

Timescale, Milestones and KPIs

M6	3	Simone Riggi - INAF	**TAR3.9 — Astronomical images segmentation with Machine Learning: Produce an updated version of caesar-mrcnn source finder based on TensorFlow v2** Produce an updated version of caesar-mrcnn source finder based on TensorFlow v2.	caesar-mrcnn software update	Source code repository url: https://github.com/SKA-INAF/caesar-mrcnn-tf2	27/8/2023
M6	3	Thomas Cecconello - INAF	**TAR3.10 — Self-Supervised techniques in Radioastronomy: first code implementation and dataset collection.** Produced a codebase to test self supervised techniques, in particular BYOL	First codebase released	https://github.com/dr4thmos/byol	27/8/2023
M6	3	Thomas Cecconello - INAF	Produced datasets useful for self supervised learning in radio continuum domain	Dataset v1	https://docs.google.com/spreadsheets/d/1J4ycn8iy1TsXzHR0OiQDAEF4upec5OeSLf5Kg4E-DDU/edit?usp=share_link	27/8/2023

Timescale, Milestones and KPIs

M7	3	Simone Riggi - INAF (3.8)	Run caesar-mrcnn with alternative backbones (ResNet18) pre-trained on unlabelled data with self-supervised contrastive learning	caesar-mrcnn backbone pre-training runs	Contributed presentation at the ADASS 2023 conference: https://adass2023.lpl.arizona.edu/events/c402	27/10/2023
M7	3	Thomas Cecconello - INAF (3.9)	Added two datasets to the collection. One is Banner, a more curated dataset with sources in the middle (subset of hulk). The second one is RGZ DR1 dataset, retrieved from a crowd labeling campain on radio galaxy zoo.	Dataset v2	https://docs.google.com/spreadsheets/d/1MZ9f0-pHTYm6FGMzNtL7dT6w7PXyi9RlvqLIeAtxDWM/edit?usp=share_link	27/10/2023
M7	3	Thomas Cecconello - INAF (3.9)	Forked solo-learn library that provides SOTA methods of SSL. Implementation of custom classes for radio images and custom augmentations.	Repository commit	https://github.com/dr4thmos/solo-learn-radio/commit/04cfaf54b976f63902c501fa8343b8d9d9c52007	nov 29, 23

Timescale, Milestones and KPIs

M8	3	Simone Riggi - INAF (3.8)	Presentation at CERAML workshop	Link to the conference	https://www.um.edu.mt/newspoint/events/um/2024/03/workshop-centre-of-excellence-in-radio-astronomy-and-machine-learning-ceraml	27/03/2023
M8	3	Thomas Cecconello - INAF (3.9)	Retrieved 3 more datasets for the downstream classification tasks, namely: VLASS, FRG, MiraBest	Dataset v3	https://docs.google.com/spreadsheets/d/1tekXnxrBA3scV7hSIbjm1j-qGSEuwQsc36I-XUpMibs/edit?usp=sharing	27/03/2023
M8	3	Thomas Cecconello - INAF (3.9)	• Added more methods for evaluations: k-fold cross validation in different fashions, fixed or random. • Added automatic generation of evaluation experiment for benchmarking. • Added top 2 accuracy. • Added all4one method	Repository commits	https://github.com/dr4thmos/solo-learn-radio/commits/master/	27/04/2023
M8	3	Thomas Cecconello - INAF (3.9)	Presentation at CERAML workshop	Link to the conference	https://www.um.edu.mt/newspoint/events/um/2024/03/workshop-centre-of-excellence-in-radio-astronomy-and-machine-learning-ceraml	27/03/2023