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

CANDELA – ITHACA s.r.l.

standard **CAN**dle-based **D**istance **E**stimation with **L**earning **A**lgorithms

Andrea Lessio, Vanina Fissore, Virginia Ajani, Paolo Viviani, Martina Giovalli, Beatrice Bucciarelli, Deborah Busonero

Spoke 3 II Technical Workshop, Bologna Dec 17 -19, 2024

Scientific Rationale

- ESA Satellite **Gaia** has delivered a massive amount of data (**DR3** ~ 10 TB)
- Leverage advantages of **machine learning/deep learning techniques** to extract useful information encoded in the data
- **Goal**: development of algorithms and models using Machine Learning/Deep Learning techniques for **estimating astronomical parameters** (e.g. parallax, distance) for the analysis of data from the Gaia space satellite for different types of distance indicators (**RR Lyrae, Cepheids**) and data (**catalogs**, **photometric series, astronomical plates**)
- **ITHACA s.r.l** has expertise in big data processing, image processing and machine learning techniques

Technical Objectives, Methodologies and Solutions (1/3)

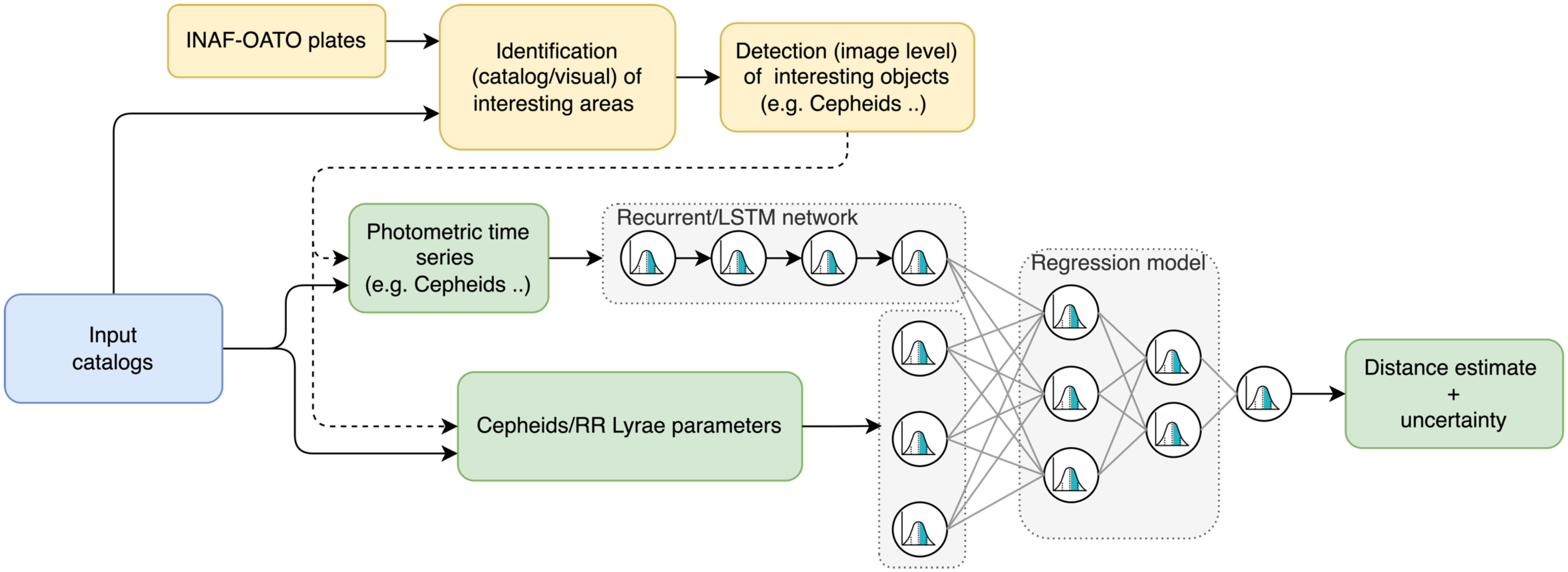
O1	Development of an ML model using analysis of photometric time series and stellar parameters, for inference of distance of Cepheid-type standard candles. Validation on a reference dataset provided by INAF-OATO.
Methodology	Integration of time series based models (e.g. LSTM) with models for tabular data (e.g. MLP) to correlate information from photometric series with astrometric and astrophysical parameters from catalogs (e.g. Gaia DR3) to eventually infer the astrometric distance of standard candles.
O2	Study of the propagation of uncertainties for the class of models (i.e., deep neural networks, recurrent neural networks) of interest, with the aim of providing an accurate estimate of the uncertainty on the predicted distance.
Methodology	Starting from model class identified in **O1**, integration of three source of uncertainty: ML model uncertainty, uncertainty on catalogue paramteres, uncertainty on ground truth distance. A possible strategy could leverage variational networks to incorporate such uncertainties.
O3	Extension, adaptation of the model developed for Cepheids to standard candle type RR Lyrae. Validation with a reference dataset provided by INAF-OATO.
Methodology	Extend model developed in **O1** to different type of standard candles: in pratice adapt and re-train the algoritm to take as input RR Lyrae parameters and photometric time series.

Technical Objectives, Methodologies and Solutions (2/3)

O4	Identification of areas of interest using existing catalogs and visual inspection from astronomical plates `.fits` images provided by INAF-OATO.
Methodology	Check correspondance of the objects present in the digitalized astronomical plates both by comparison with existing catalogs (e.g. GAIA, OGLE) at coordinate level and by visual inspection with the scientific support of INAF-OATO
O5	Detection of interesting objects (e.g. Cepheids, RR Lyrae) in such areas of interest to futher enrich the standard candles catalogs with complementary information and generalize the developed algorithm on a different input dataset
Methodology	Using existing softwares (e.g. SExtractor, Astrometry.net) or developing ML-based detection algoritm perform detection at the image level of Cepheids and RR Lyrae and interesting objects in astronomical plates images. Include the available sources in training sets for models developed in **O1** and **O3** to enrich the dataset and generalise the model, with the scientific support of INAF-OATO.

Technical Objectives, Methodologies and Solutions (3/3)

- **Input**: Gaia DR3, OGLE catalog, astronomical plates from INAF-OATO



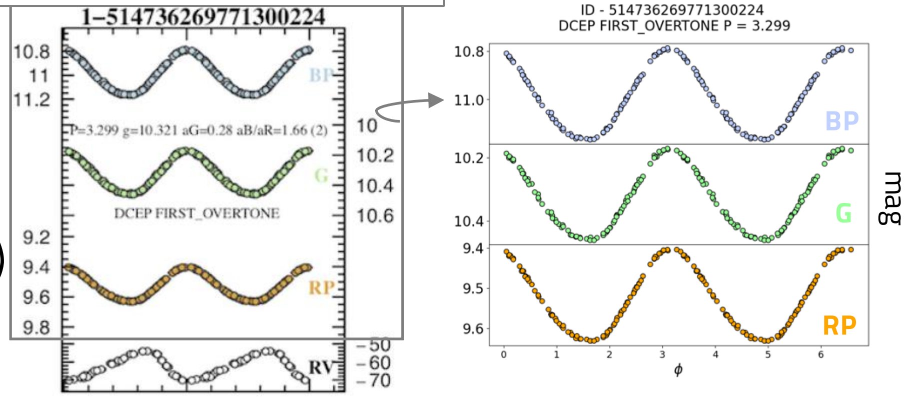
- **Output**: generalized distance estimation with learning algorithms

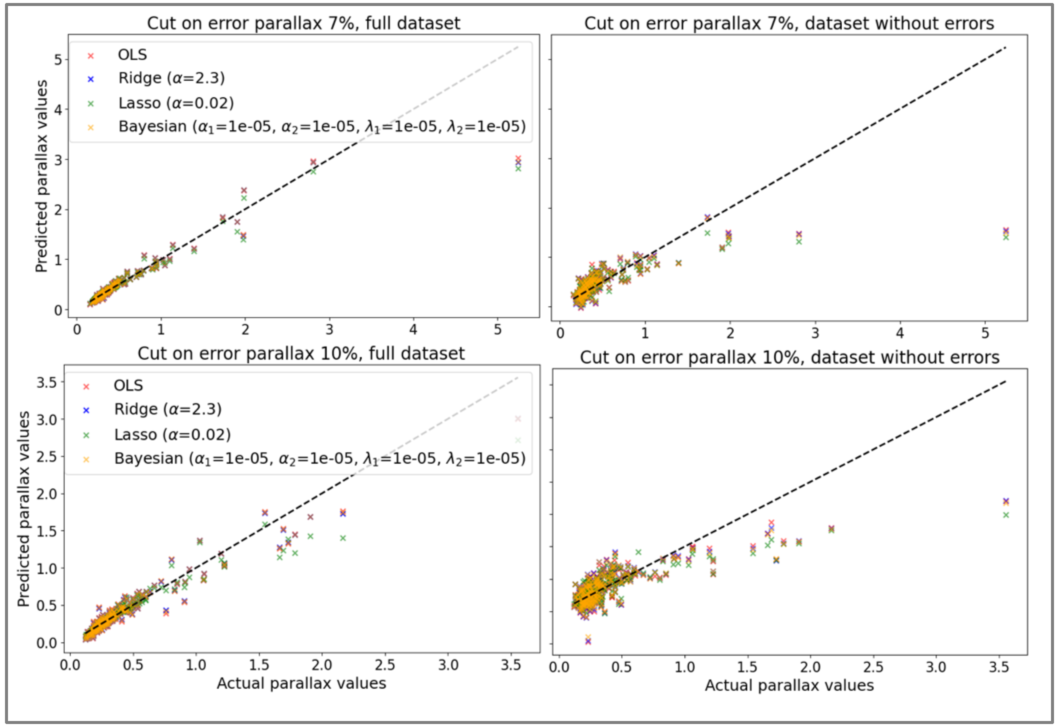
Main results achieved so far (1/2)

WP1:

- Validation of Cepheids photometric dataset to enrich input
- Creation of reduced and complete datasets with cuts on relative error on parallax
- Comparison of several classical ML algorithms to set baseline (ongoing work on MLP)

From Gaia DR3 documentation, **V. Ripepi et al.**





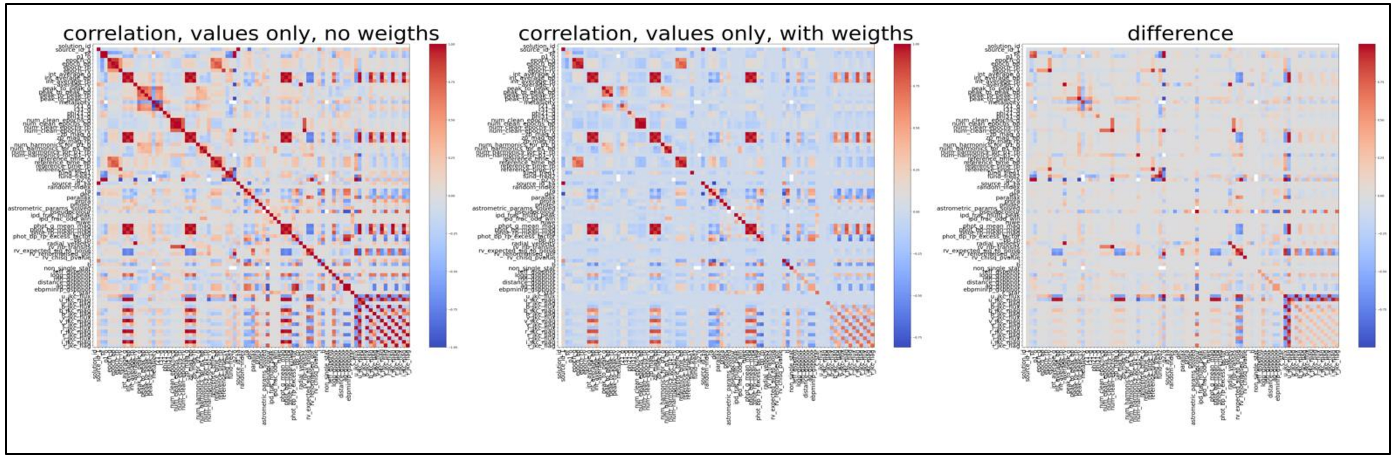
Model	7% (Full Dataset)		10% (Full Dataset)		20% (Full Dataset)	
	MSE	R^2	MSE	R^2	MSE	R^2
OLS	0.037412	0.867919	0.010137	0.929578	0.004594	0.891334
Ridge	0.040303	0.857714	0.009575	0.933484	0.004449	0.894753
Lasso	0.043161	0.847625	0.013042	0.909399	0.003744	0.911443
Bayesian	0.039743	0.859691	0.009817	0.931801	0.004496	0.893657

Model	7% (Without Errors)		10% (Without Errors)		20% (Without Errors)	
	MSE	R^2	MSE	R^2	MSE	R^2
OLS	0.131349	0.536286	0.056317	0.608771	0.022739	0.462134
Ridge	0.131010	0.537480	0.056753	0.605742	0.022755	0.461748
Lasso	0.140025	0.505654	0.056523	0.607342	0.020106	0.524409
Bayesian	0.133096	0.530119	0.056016	0.610867	0.022582	0.465854

Main results achieved so far (2/2)

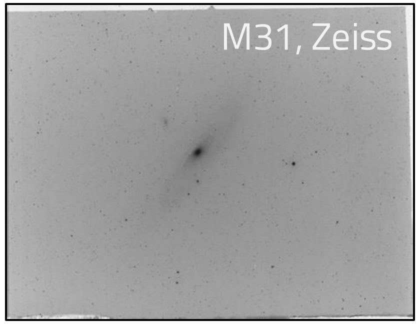
WP2:

- Study of imputation of parameters with missing data (e.g. metallicity, more than 9000 missing)
- Currently investigating correlations among parameters from Cepheids catalog, feature importance



WP4:

- Set-up methodology to identify INAF-OATO historical archive plates with objects of interests (Cepheids, RR Lyrae)



Example plate image from: https://astroarchive.oato.inaf.it/Plates/

Current status and next steps

- **Status with respect to the timescale and milestones:**

Month	1	2	3	4	5	6	7	8	9	10	11	12
WP1 - Model with Cepheids catalog	x	x	x	x								
WP2 - Study of uncertainties propagation				x	x	x	x	x	x			
WP3 - Model with RR Lyrae catalog							x	x	x			
WP4 - Identify INAF-OATO plates of interest				x	x	x						
WP5 - Object detection on plates, enrich input								x	x	x	x	x
				MS1		MS2			MS3, MS4			MS5

- **Next steps:**

WP1: identify first candidate model based on obtained metrics and performance, validate output

WP2: continue investigation of uncertainty propagation due to data imputation and model error

WP3: will start at Month 7 → extend model to RR Lyrae catalog

WP4: starting from plates coordinates and FOV, look for areas of interest matching Gaia catalog

WP5: will start at Month 8 → detect interesting objects in plates found in WP4, potentially enrich input datasets

Thank you for your attention!

Spoke 3 II Technical Workshop, Bologna Dec 17 -19, 2024