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

Machine Learning and Deep Learning algorithms for Gaia mission data analysis

Lorenzo Monti, Tatiana Muraveva

INAF - Osservatorio di Astrofisica e Scienza dello Spazio di Bologna

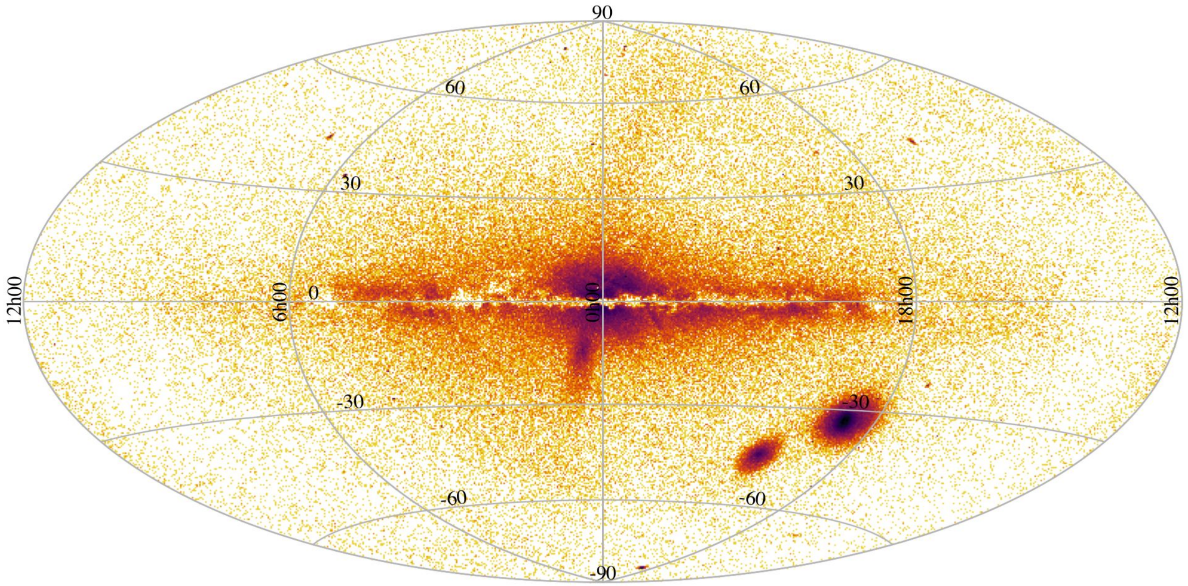
Spoke 3 General Meeting, Elba 5-9 / 05, 2024

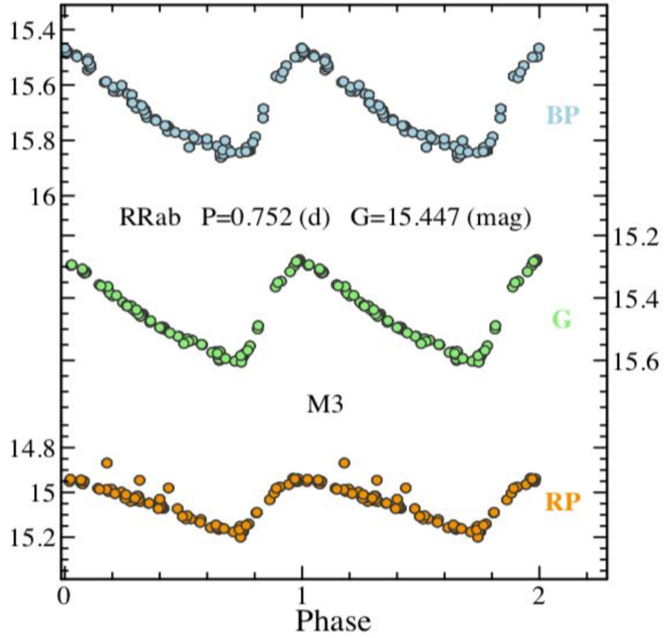
Scientific Rationale

➔ RR Lyrae stars are periodic (Period < 1 day), pulsating, variable stars that play a crucial role in stellar astrophysics.

➔ There is a correlation between RRL's light curves and their metallicities ([Fe/H]).

➔ Gaia Data Release 3 provides a catalogue of 270 905 RRLs along with their time-series photometry.





Project Main Goal: Derive metallicities of RR Lyrae stars from their time-series photometry data using Machine Learning/Deep Learning algorithms.

Technical Objectives, Methodologies and Solutions

Time-series Extrinsic Regression TSER(1) is a *regression task* that learns the mapping from time series data to a scalar value. That *task* depend on the whole series, rather than depending more on recent than past values such as time-series forecasting (**TSF**).

The difference between *time series classification* (**TSC**) and **TSER** is that TSC maps a time series to a finite set of discrete labels while TSER predicts a continuous value from the time series.

As described in *Tan, et al.* (1), a **TSER model** is a function $\mathcal{T} \rightarrow \mathcal{R}$, where $\mathcal{T}$ is a class of time series and $\mathcal{R}$ a class of scalar values. **TSER** seeks to learn a regression model from a dataset $\mathcal{D} = \{(t_1, r_1), \ldots, (t_n, r_n)\}$, where t_i is a time series and r_i is a continuous scalar value.

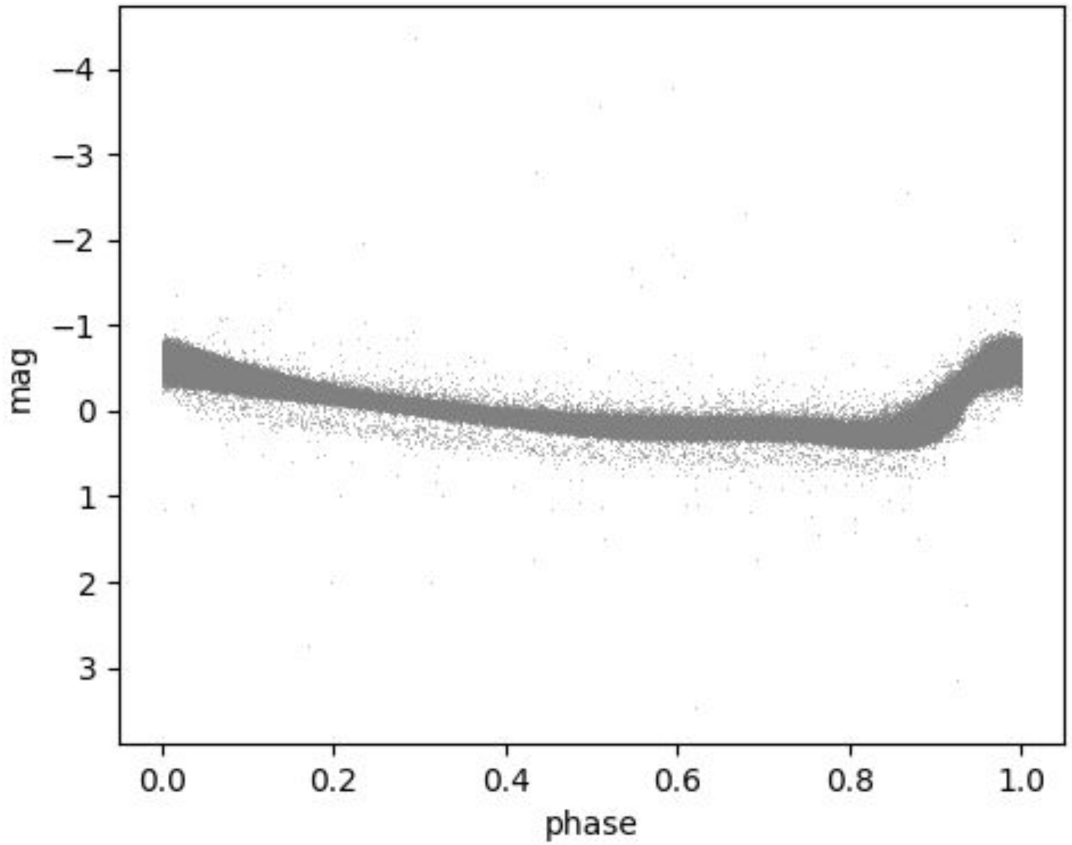
1. Tan, Chang Wei, et al. "Time series extrinsic regression: Predicting numeric values from time series data." *Data Mining and Knowledge Discovery* 35 (2021): 1032-1060.

Technical Objectives, Methodologies and Solutions

Dataset preparation

As regarding the time-series photometry dataset we selected a set of **6696 RRab stars** based on:

- err[Fe/H] < 0.4 dex
- peak-to-peak amplitude < 1.4 mag
- Number of epochs > 50
- ϕ_{31} error < 0.10



Technical Objectives, Methodologies and Solutions

Data pre-processing
01

For the predictive modeling of the [Fe/H] from the light curves, we use the following two-dimensional sequences as input variables:

$$X^{<t>} = \begin{cases} m^{<t>} - < m > \\ Ph * P \end{cases} \qquad t = 1, ..., N_{ep}$$

where $m^{\langle t \rangle}$ is the magnitude of the light curve, $<m>$ is the mean magnitude, *Ph* is the phase and *P* the period. N_{ep} is the number of epochs.

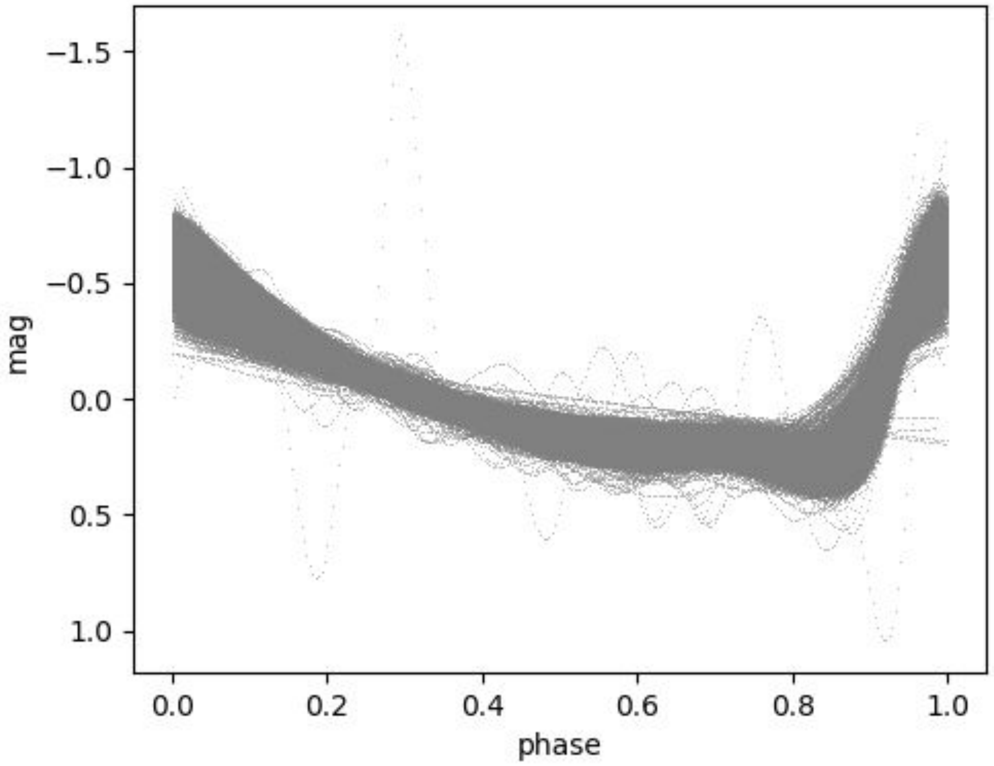
Technical Objectives, Methodologies and Solutions

Data pre-processing
02

After that, **(1)** we applied the *spline smoothing* method.

The method is applied **(2)** to minimize fluctuations, noise, outliers and obtain the same number of points for each light curve (264).

(3) Finally the pre-processed catalog is stored. The script *pre-processing.py* is contained within the micro-library(1) written to obtain the estimation model.



So, the final input tensor for ML/DL methods have a shape of [**6696, 264, 2**] ⟶ **[** batch size, time steps, features**]**.

1. Micro-library to estimate metallicity from RR-Lyrae phototometric light curves time series through machine learning and deep learning models. Link:https://github.com/LorenzoMonti/metallicity_rrls

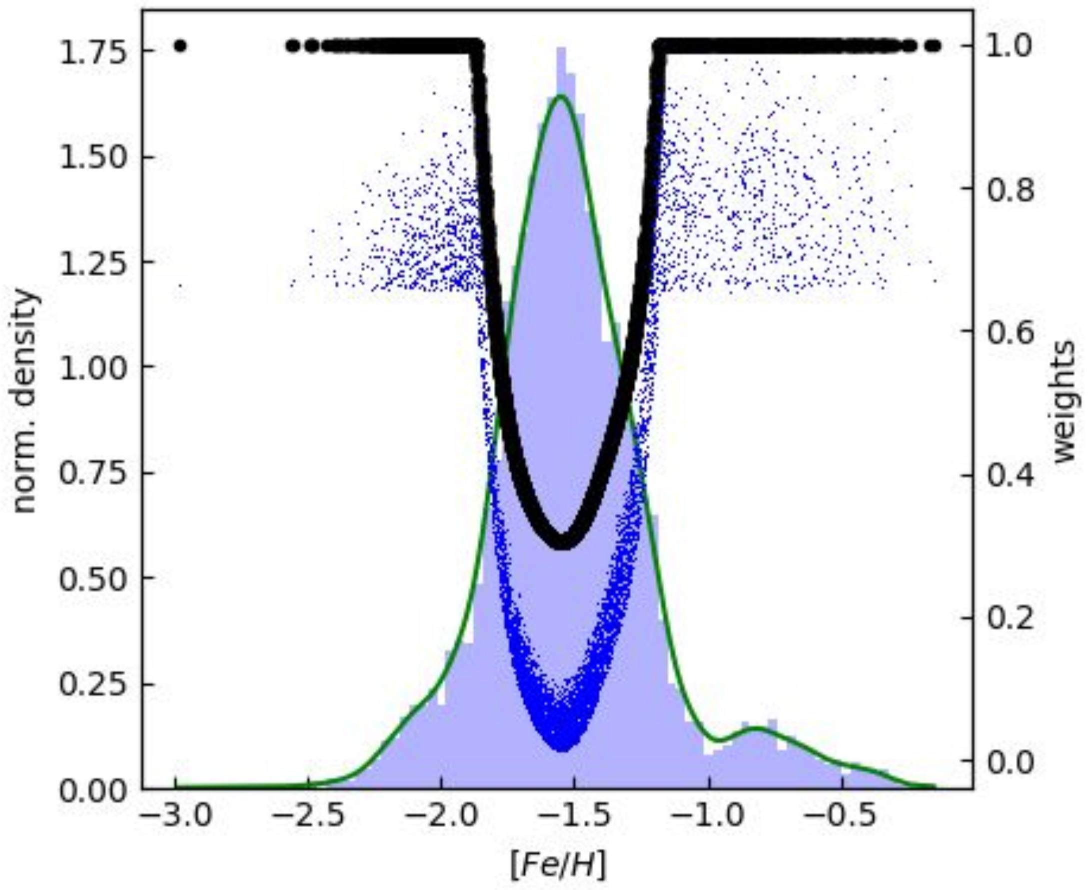
Technical Objectives, Methodologies and Solutions

Data pre-processing
03

Sample weights for metallicity distribution

Based on step 2, we computed Gaussian kernel density estimates of the [Fe/H] distributions.

Evaluated them for every object in the datasets, and assigned a density weight w_d to each data point by taking the inverse of the estimated normalized density.



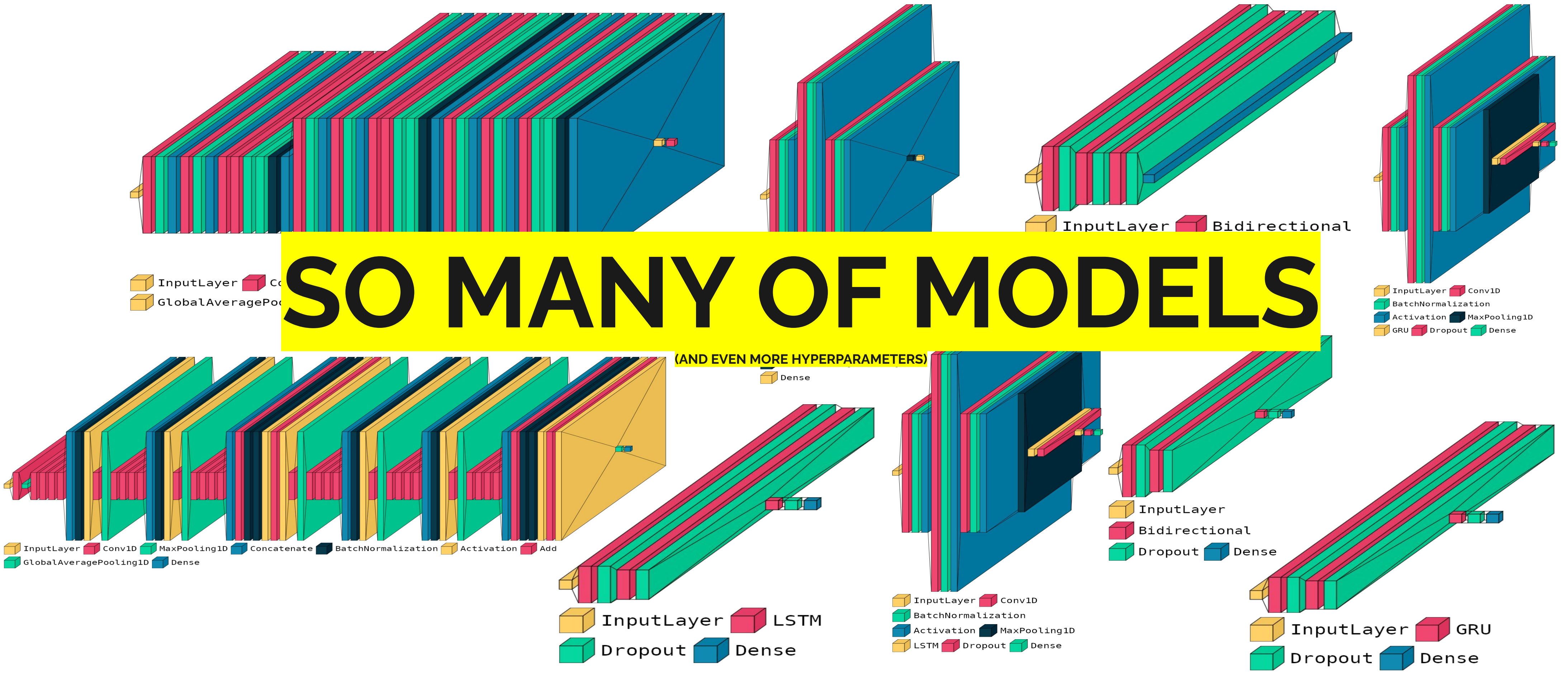
Technical Objectives, Methodologies and Solutions

Several both Machine Learning and Deep Learning models have been created. As regards Deep Learning models, **Convolutional** models, **Recurrent** models and **Mixed architectures** among these were taken into consideration.

Machine Learning Models	Convolutional Neural Networks	Recurrent Neural Networks	Mixed architectures
- Random forest - Support Vector Regressor	- FCN - Resnet - InceptionTime	- GRU - BiGRU - LSTM - BiLSTM	- ConvGRU - ConvLSTM

Each model was trained on **3 different datasets**: (i) *raw dataset without spline*, (ii) *raw dataset with spline* and (iii) *preprocessed dataset with spline*. This is to verify the actual contribution of the preprocessing. The results obtained are the average resulting from the (stratified) **K-fold cross validation** (n_splits=5).

Technical Objectives, Methodologies and <u>Solutions</u>

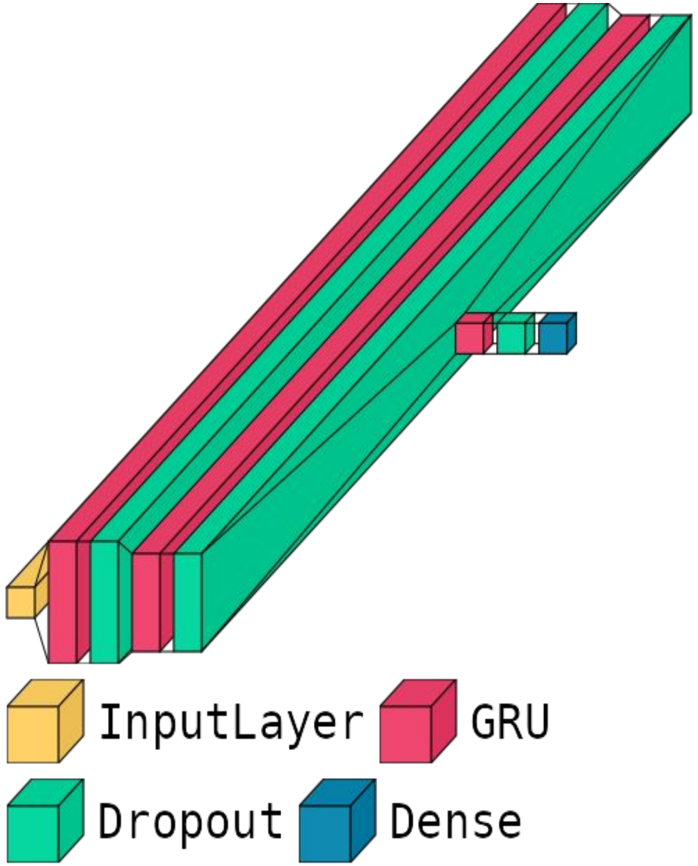


SO MANY OF MODELS

(AND EVEN MORE HYPERPARAMETERS)

Technical Objectives, Methodologies and Solutions

Architecture based on GRU



Layer (type)	Output Shape	Param #
input_1 (InputLayer)	[(None, 264, 2)]	0
gru (GRU)	(None, 264, 20)	1440
dropout (Dropout)	(None, 264, 20)	0
gru_1 (GRU)	(None, 264, 16)	1824
dropout_1 (Dropout)	(None, 264, 16)	0
gru_2 (GRU)	(None, 8)	624
dropout_2 (Dropout)	(None, 8)	0
dense (Dense)	(None, 1)	9

Accomplished Work, Results

Steps done

Catalogue of RR Lyrae stars and **Metallicity** validation
LSTM implementation but also BiLSTM, GRU, BiGRU, convolutional models and mixed architectures.

May 2024

Oct 2023

Steps to do

Oct 2024

Transformers architecture implementation.

The best results use the dataset with smoothing splines and averaging in magnitude. Among these, RNNs are the type of network with the best results and in particular **GRU** and **BiGRU** have obtained practically the same results.

R^2	RMSE	MAE
0.9451	0.0739	0.0547
on validation set and 0.9472 on training set	on validation set and 0.0724 on training set	on validation set and 0.0537 on training set

Accomplished Work, Results

	GRU		BiGRU	
	training	validation	training	validation
r2	0,947	0,9449	**0,9472**	**0,9451**
wrmse	0,0723	0,0736	**0,0721**	**0,0735**
wmae	0,0545	0,0551	**0,0534**	**0,0544**
rmse	0,0727	0,074	**0,0724**	**0,0739**
mae	0,0548	0,0554	**0,0537**	**0,0547**

catalog_raw_spline

	LSTM		BiLSTM		GRU		BiGRU		ConvLSTM		ConvGRU							
Metrics	training	validation	training	validation	training	validation	training	validation	training	validation	training	validation						
r2	0,9303	0,9275	0,9328	0,932	0,9396	0,936	0,942	0,9375	0,9367	0,9236	0,9372	0,9306	0,9286	0,925	0,9462	0,9314	**0,9489**	**0,9378**
w rmse	0,0828	0,0845	0,0813	0,0818	0,0771	0,0794	0,0755	0,0784	0,079	0,0868	0,0786	0,0827	0,0838	0,0859	0,0728	0,0822	**0,0709**	**0,0783**
w mae	0,0612	0,0622	0,0601	0,0604	0,0564	0,0577	0,0568	0,0584	0,0586	0,0637	0,0583	0,0609	0,0612	0,0616	0,0531	0,059	**0,0522**	**0,0571**
rmse	0,0833	0,085	0,0818	0,0822	0,0776	0,0798	0,076	0,0788	0,0795	0,0872	0,0792	0,0833	0,0842	0,0863	0,0731	0,0824	**0,0713**	**0,0786**
mae	0,0616	0,0626	0,0605	0,0607	0,0568	0,058	0,0572	0,0588	0,0591	0,0641	0,0588	0,0614	0,0615	0,0619	0,0534	0,0592	**0,0525**	**0,0574**
note																		
CLASSIFICA	resnet		gru		bigru		Inception		bilstm		convgru		lstm		fcn		convlstm	

catalog_mean_spline

	LSTM		BiLSTM		GRU		BiGRU		ConvLSTM		ConvGRU		FCN		Inception		RESNET	
Metrics	training	validation	training	validation	training	validation	training	validation	training	validation	training	validation	training	validation	training	validation	training	validation
r2	0,9353	0,9306	0,942	0,9375	0,947	0,9449	**0,9472**	**0,9451**	0,9485	0,9357	0,9433	0,9361	0,9402	0,9368	0,9546	0,9442	0,9507	0,9388
w rmse	0,0798	0,0827	0,0755	0,0784	0,0723	0,0736	**0,0721**	**0,0735**	0,0712	0,0796	0,0747	0,0793	0,0767	0,0789	0,0669	0,0742	0,0697	0,0776
w mae	0,0605	0,062	0,0568	0,0584	0,0545	0,0551	**0,0534**	**0,0544**	0,0539	0,0595	0,057	0,0597	0,0571	0,058	0,0505	0,0554	0,0524	0,0576
rmse	0,0803	0,0831	0,076	0,0788	0,0727	0,074	**0,0724**	**0,0739**	0,0716	0,08	0,0751	0,0797	0,077	0,0791	0,0672	0,0745	0,07	0,0779
mae	0,0609	0,0624	0,0572	0,0588	0,0548	0,0554	**0,0537**	**0,0547**	0,0543	0,0599	0,0573	0,06	0,0574	0,0583	0,0508	0,0556	0,0526	0,0579
note																		
CLASSIFICA	bigru		gru		inception		bilstm		resnet		fcn		convgru		convlstm		lstm	

catalog_raw_nospline

	LSTM		BiLSTM		GRU		BiGRU		ConvLSTM		ConvGRU		FCN		Inception		RESNET	
Metrics	training	validation	training	validation	training	validation	training	validation	training	validation	training	validation	training	validation	training	validation	training	validation
r2	0,85	0,82	0,8939	0,8574	**0,9**	**0,876**		0,217	err	err	0,937	0,7148	0,8898	0,7188	0,966	0,7888	0,954	0,7863
w rmse			0,1022	0,1185		**0,11**		0,277	err	err	0,0787	0,1676	0,1042	0,1664	0,0579	0,1442	0,0673	0,1451
w mae			0,0789	0,0918		**0,847**		0,225	err	err	0,0581	0,1315	0,0786	0,1297	0,037	0,1135	0,0443	0,1132
rmse	0,115	0,126	0,1027	0,119		**0,11**		0,275	err	err	0,0794	0,168	0,1049	0,1669	0,0584	0,1447	0,0676	0,1455
mae			0,0794	0,0923		**0,851**		0,223	err	err	0,0585	0,1319	0,0792	0,1302	0,0374	0,1139	0,0446	0,1135
note									diverge-> r2 basso									
CLASSIFICA	gru		bilstm		lstm		inception		resnet		convgru		fcn		convlstm		bigru	

The table presents the results, in terms of R^2, *wrmse, wmae, rmse and mae*, of all ML/DL models and for each datasets.

In the focus, the results of the models with best performance, **GRU** and **BiGRU**, are shown.

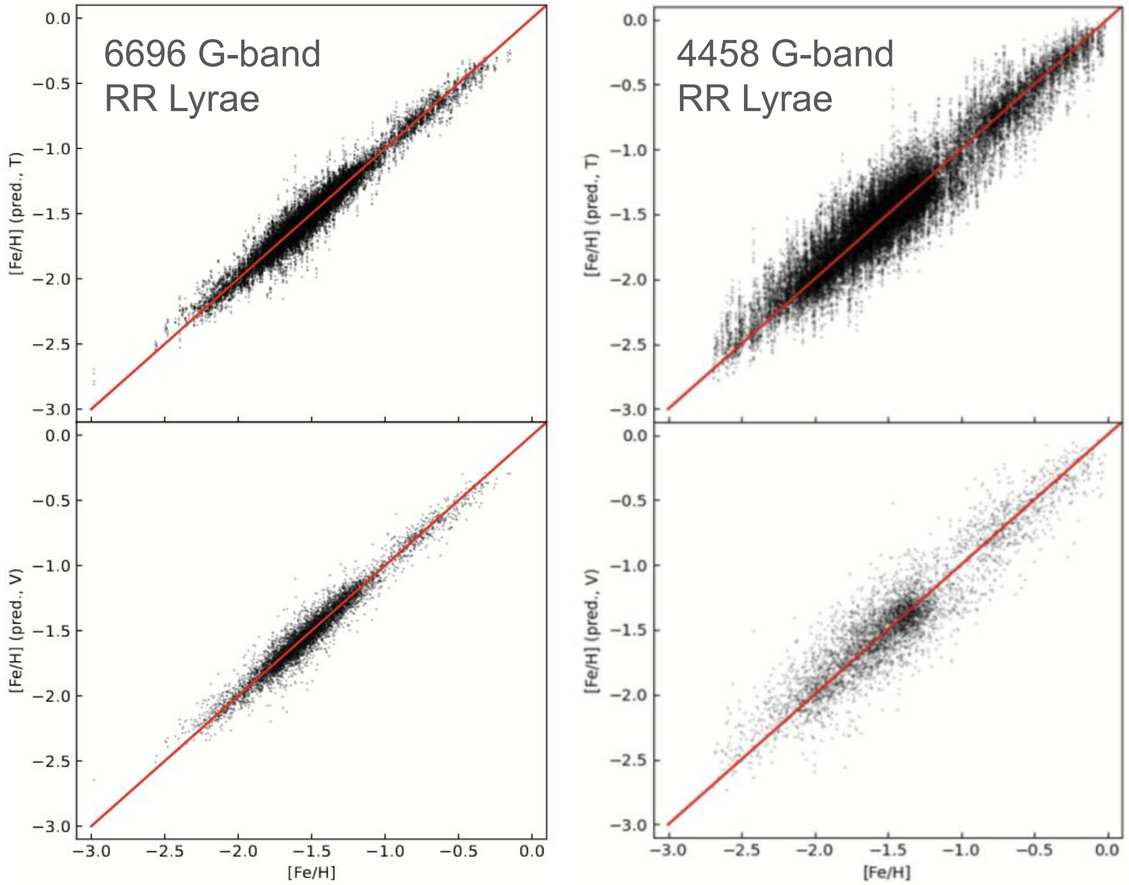
Accomplished Work, Results

We compared our results with the best result in scientific literature.

Our results	GRU	
	training	validation
r2	0,947	0,9449
wRMSE	0,0723	0,0736
wMAE	0,0545	0,0551
RMSE	0,0727	0,074
MAE	0,0548	0,0554

Dekany's results	BiLSTM	
	training	validation
r2	0,96	0,93
wRMSE	0,1	0,13
wMAE	0,07	0,1
RMSE	0,15	0,18
MAE	0,12	0,13

The plot on the left shows **our metallicity prediction results** while the one on the right shows **Dekany's metallicity prediction** results(1).



1. Dékány, István, and Eva K. Grebel. "Photometric Metallicity Prediction of Fundamental-mode RR Lyrae Stars in the Gaia Optical and K s Infrared Wave Bands by Deep Learning." *The Astrophysical Journal Supplement Series* 261.2 (2022).

Timescale, Milestones and KPIs

DEC 2023 – OCT 2024

Building Neural Networks such as **LSTM** and **Transformers architecture** in order to estimate metallicity of RR Lyrae stars from time-series photometry based on the catalogue produced on the step (1).

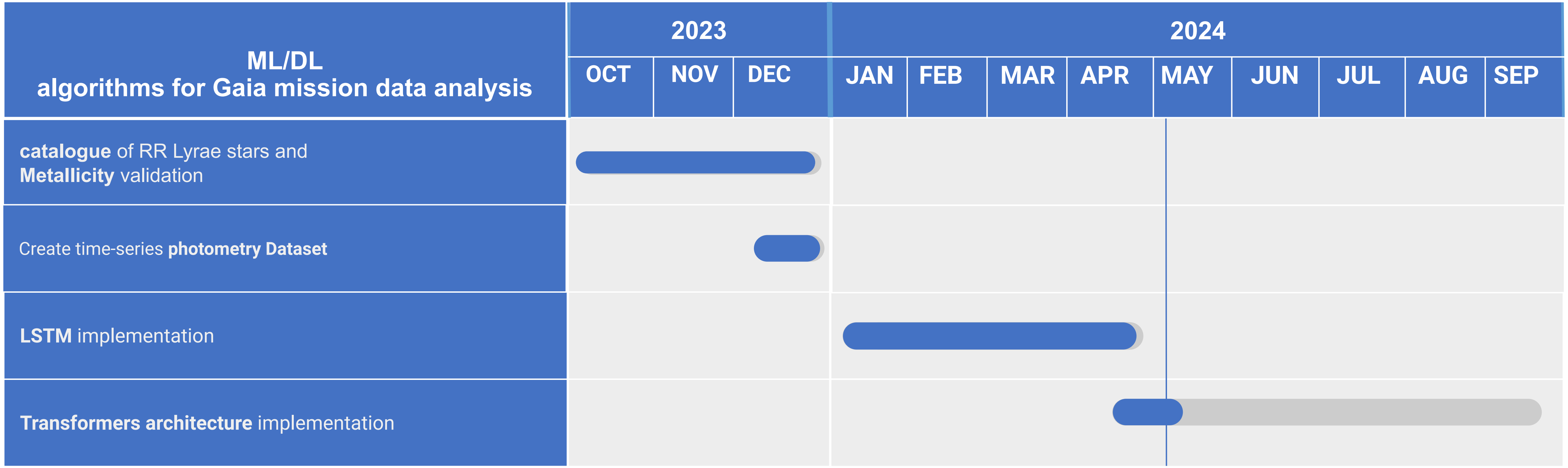
ML/DL algorithms for Gaia mission data analysis	2023			2024								
	OCT	NOV	DEC	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP
catalogue of RR Lyrae stars and **Metallicity** validation												
Create time-series **photometry Dataset**												
LSTM implementation												
Transformers architecture implementation												

Slide from ICSC Italian Research Center on High-Performance Computing, Big Data and Quantum Computing in **Trieste 2023**

Timescale, Milestones and KPIs

DEC 2023 – OCT 2024

Building Neural Networks such as **LSTM** and **Transformers architecture** in order to estimate metallicity of RR Lyrae stars from time-series photometry based on the catalogue produced on the step (1).



We did everything we were expected to do. Furthermore, we worked in dataset pre-processing, have implemented other models (10) between machine learning and deep learning and an open source repository located on github. **OOP** constructs also allow for the easy implementation of new models.

Next Steps and Expected Results

Next steps:

- **Fine tuning optimization** of the hyperparameters on the models that performed in a better way.
- New architectures like **Transformers** applied to time series.
- Approaches used for Natural Language Processing such as **seq2seq** and applying them to time series.
- Work with different catalogs such as using **RR Lyrae type c**.

Expected results (and KPI):

- A **scientific paper** "Metallicity of RR Lyrae stars from the Gaia Data Release 3 catalogue exploiting Machine Learning algorithms" by Muraveva et al. submitted to MNRAS.
- A **technical paper** "Using Deep Learning to predict Photometric Metallicity of RR-Lyrae variable Stars from its light curves in Gaia G band" by Monti et al. in preparation for submission to Machine Learning and AI for Sensors Journal.
- A **scientific paper** "Utilizing Deep Learning Techniques to Analyze the Metallicity of RR Lyrae Stars in the Gaia Data Release 3 Catalogue" by Monti et al. in preparation for submission to MNRAS.
- An **open source repository** released with CD/CI pipeline and automatic release.

Thank you
for your attention

contact

Lorenzo Monti lorenzo.monti@inaf.it - **Tatiana Muraveva** tatiana.muraveva@inaf.it