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

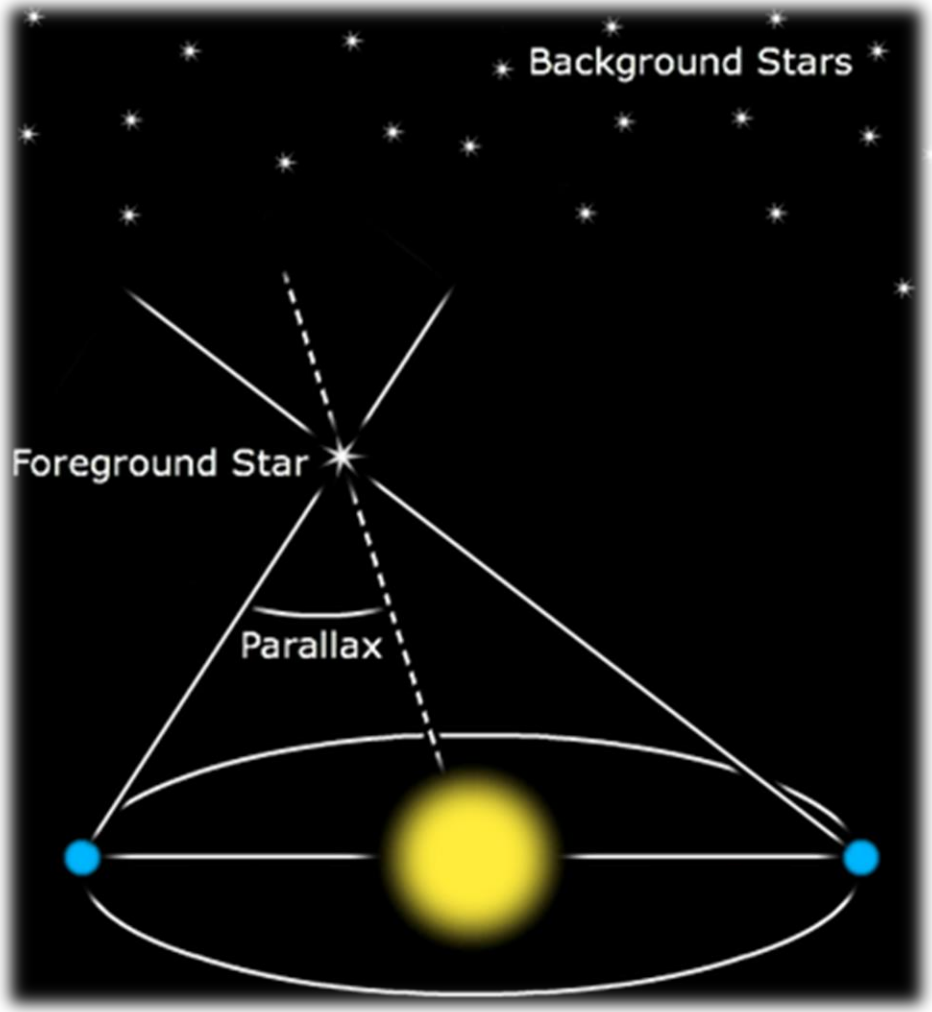
NeuroStarMap

Stefano Tortora, ALTEN

Spoke 3 III Technical Workshop, Perugia 26-29 Maggio, 2025

Project overview

- **Main goal:** development of a neural network-based algorithm to estimate astronomical distances, trained and validated on the Gaia DR3 stellar catalog.
- **Impact:** enabling astronomers to obtain more reliable estimation of stellar and cosmological distances.



Scientific coordinations

- Deborah Busonero
- Mario Gai

Team composition

- Alessio Petrone (PM)
- Luca Maina (TL)
- Stefano Tortora (TL)
- Simone Zimotti (DEV)
- Carmine Fruncillo (DEV)

Project steps

- **Dataset preparation:** collection and processing of data from GAIA satellite on three datasets concerning cepheids, RRLyrae and eclipse binaries.
- **Neural network architecture study:** analysis of potential neural network architectures.
- **Network development and implementation:** design and coding of the neural network models selected from the previous study.
- **Training of networks:** training of the networks on the datasets, monitoring performance metrics and optimizing them by tuning parameters and adapting the architecture to the specific characteristics of the datasets.
- **Validation of the networks:** preliminary testing of the neural networks on the datasets to ensure that they meet the established accuracy and efficiency criteria.
- **Final pipeline development:** creating decision mechanisms based on simple averages, weighted averages or other techniques to be defined.
- **Final system validation:** testing of the completed system for overall effectiveness, robustness and ease of use, ensuring that the final product meets design requirements.

Technologies

- Classification and regression algorithms
- Data preprocessing (scaling, encoding, normalization etc.)
- Feature selection
- Model evaluation (accuracy, confusion matrix etc.)

- Full support for building and training deep neural networks
- Management of layers, activation functions, and optimizers
- Training and validation on GPU
- Model saving, exporting, and reuse

Ideal for problems involving structured data

Ideal for working with deep models

Data Preparation – Merge and clean datasets

The Cepheid and RRLyrae datasets individually contain a relatively low number of records. It was decided to merge them to maximize the amount of data available for a single training and validation session.

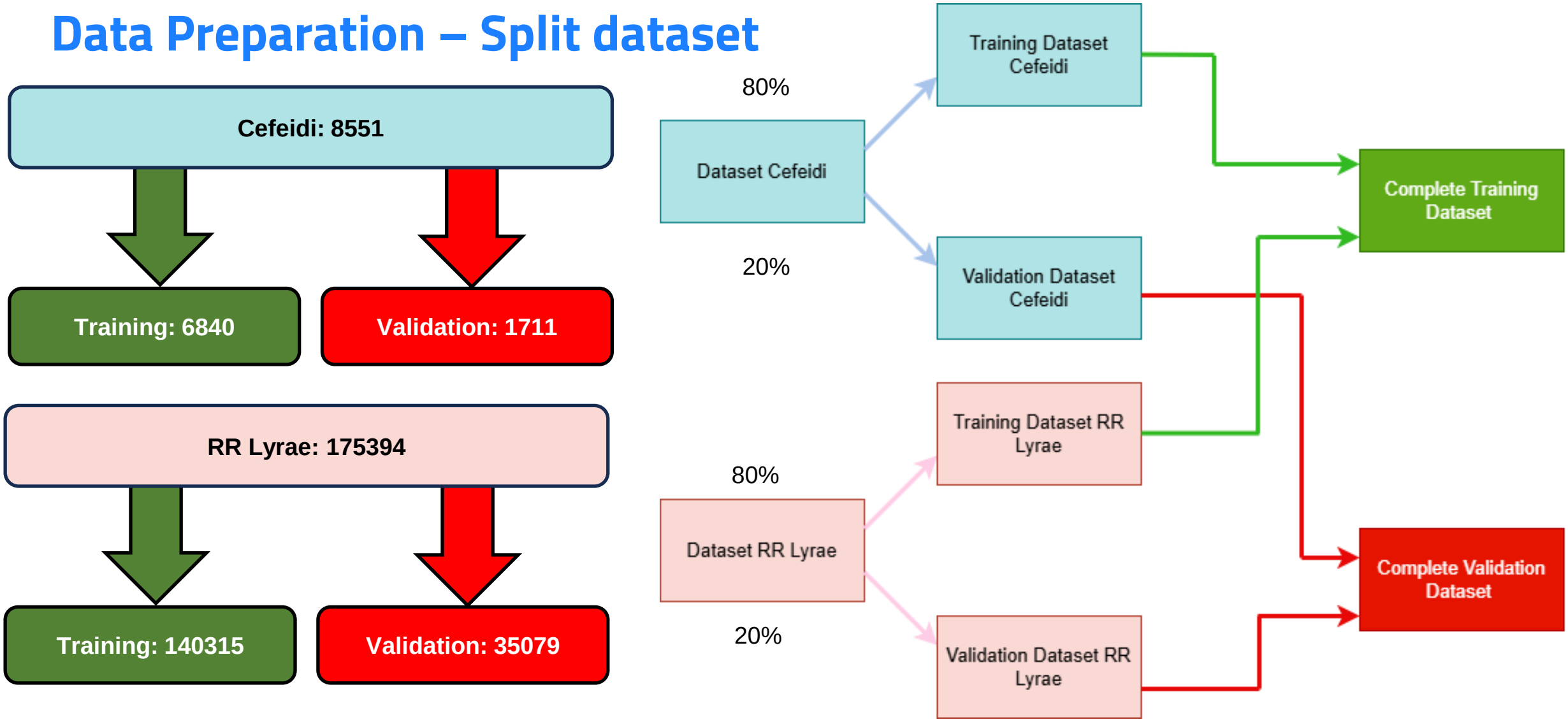
- The neural models can achieve greater generalization by learning from a wider variety of examples.

Fields common to both Cepheids and RR Lyrae stars were selected to ensure data consistency during the training and validation phases.

Name	Description
phot_g_mean_mag	G-band photometric magnitude
bp_rp	BP/RP Spectrum Photometry
peak_to_peak_g	Peak-to-peak amplitude of the G-band
metallicity	Metallicity of the star
phi21_g	Phase difference in 1 and 2 harmonic
r21_g	ratio of amplitude of the 1 and 2 harmonics
zp_mag_g	Zero point of the final G-band light curve
parallax_error	Parallax measurement error
type_best_classification	Type classification between: "DCEP", "T2CEP", "ACEP".
parallax	Parallax measurement

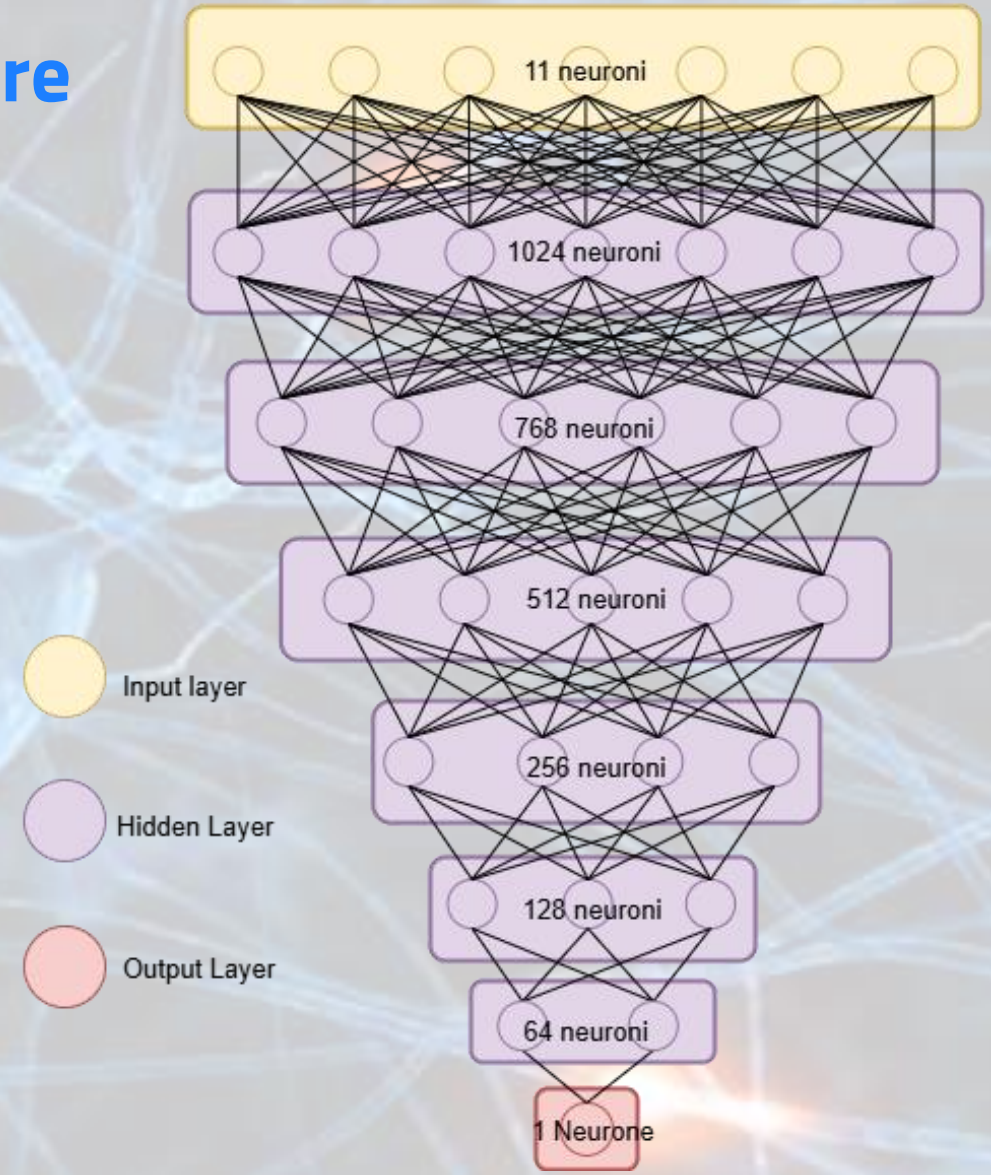
Used dataset fields

Data Preparation – Split dataset



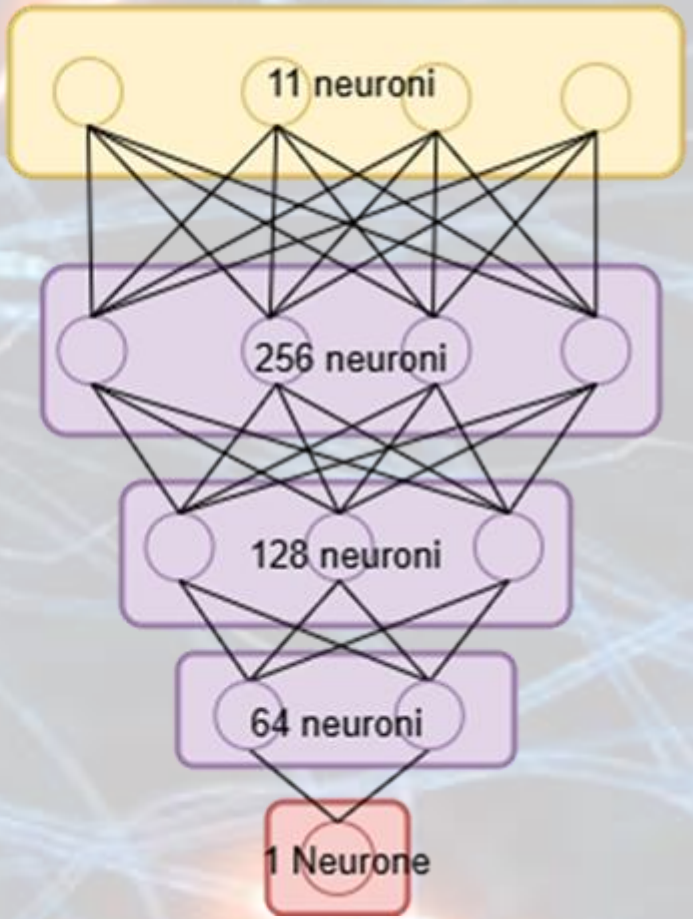
Deep Neural Network (DNN) architecture

- Very deep network (7 dense layers)
- One neuron for each feature used from the dataset in the input layer
- Dropout = 5×10^{-2}
- Weight initialization = "He Normal"
- Batch Normalization = present
- Activation function = "LeakyReLU" (alpha = 10^{-4})
- L2 Regularization = absent

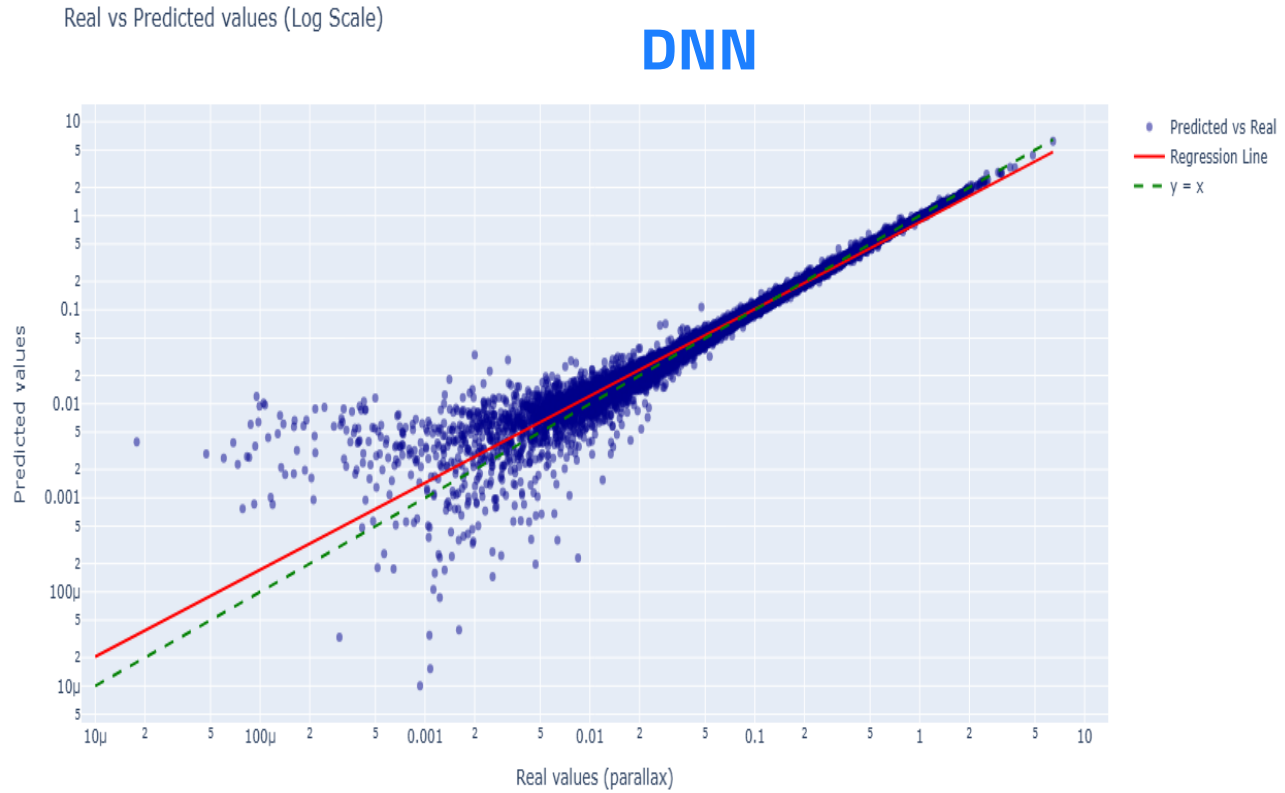


Multilayer Perceptron (MLP) architecture

- Shallow network (4 dense layers)
- One neuron for each feature used from the dataset in the input layer
- Dropout = absent
- Weight initialization = "glorot uniform"
- Batch Normalization = present
- Activation function = "LeakyReLU" (alpha = 10^{-1})
- L2 Regularization = present



DNN

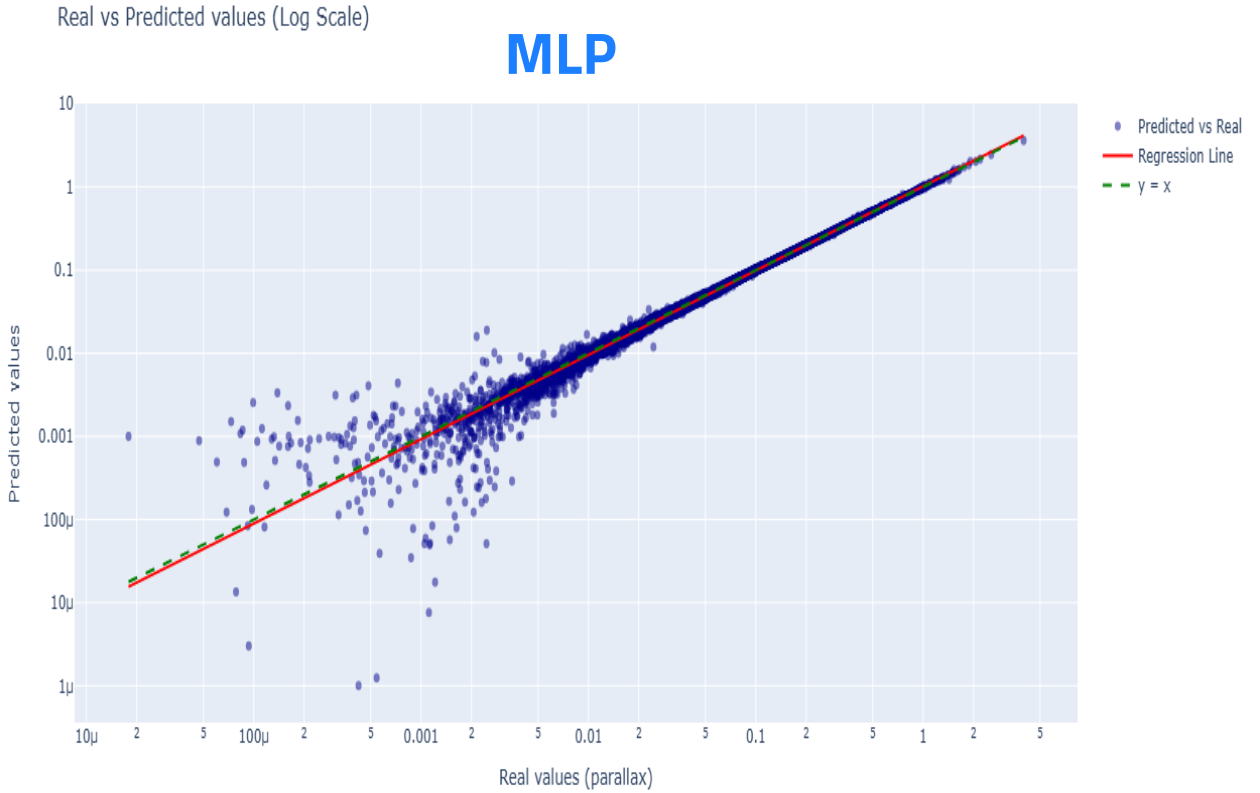


Mean Squared Error (MSE): 2.12×10^{-5}

Mean Absolute Error (MAE): 1.94×10^{-3}

Coefficient of determination (R^2): 9.989×10^{-1}

MLP



Mean Squared Error (MSE): 7.18×10^{-5}

Mean Absolute Error (MAE): 2.31×10^{-3}

Coefficient of determination (R^2): 9.958×10^{-1}

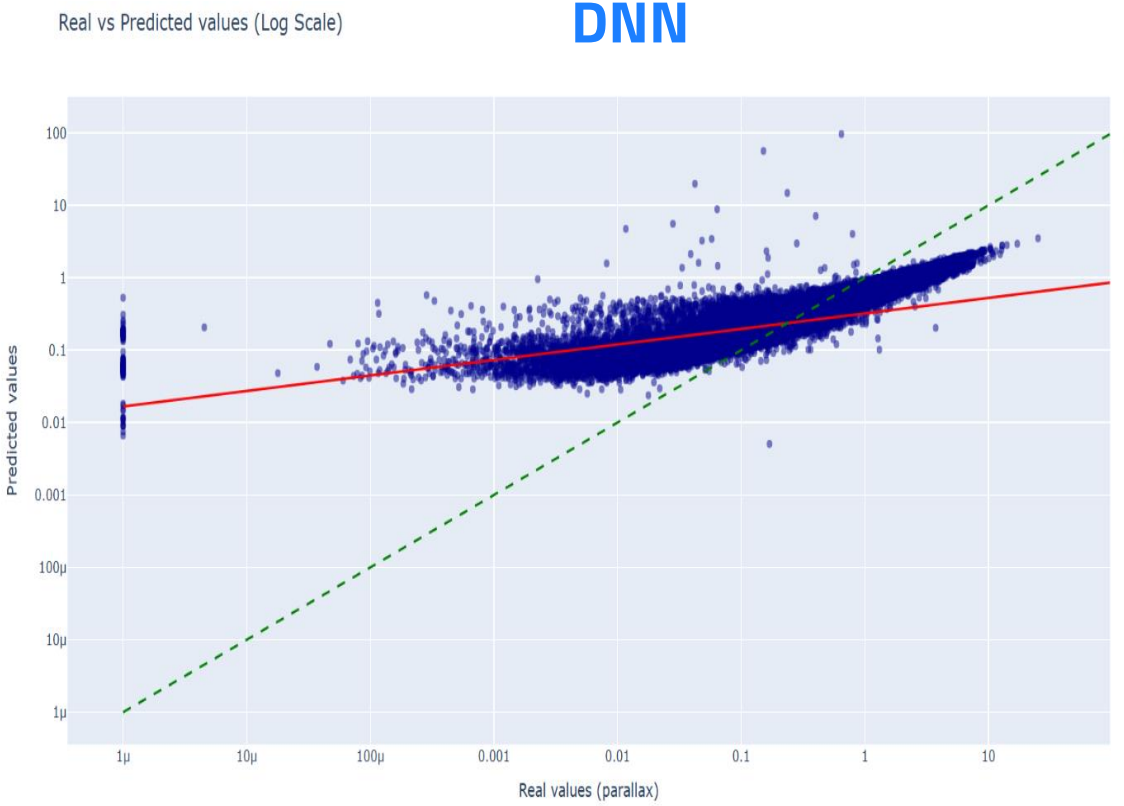
Both the DNN and MLP achieve very similar and progressively better results as the parallax values increase.

For values very close to zero the predictions become less accurate. This can be attributed to the great distance of the celestial bodies involved and to the numerical difficulties encountered by the neural networks.

Data Preparation – Clean datasets

Parallax error related fields were removed from the dataset: they may not be available in datasets that will be used for parallax prediction.

DNN

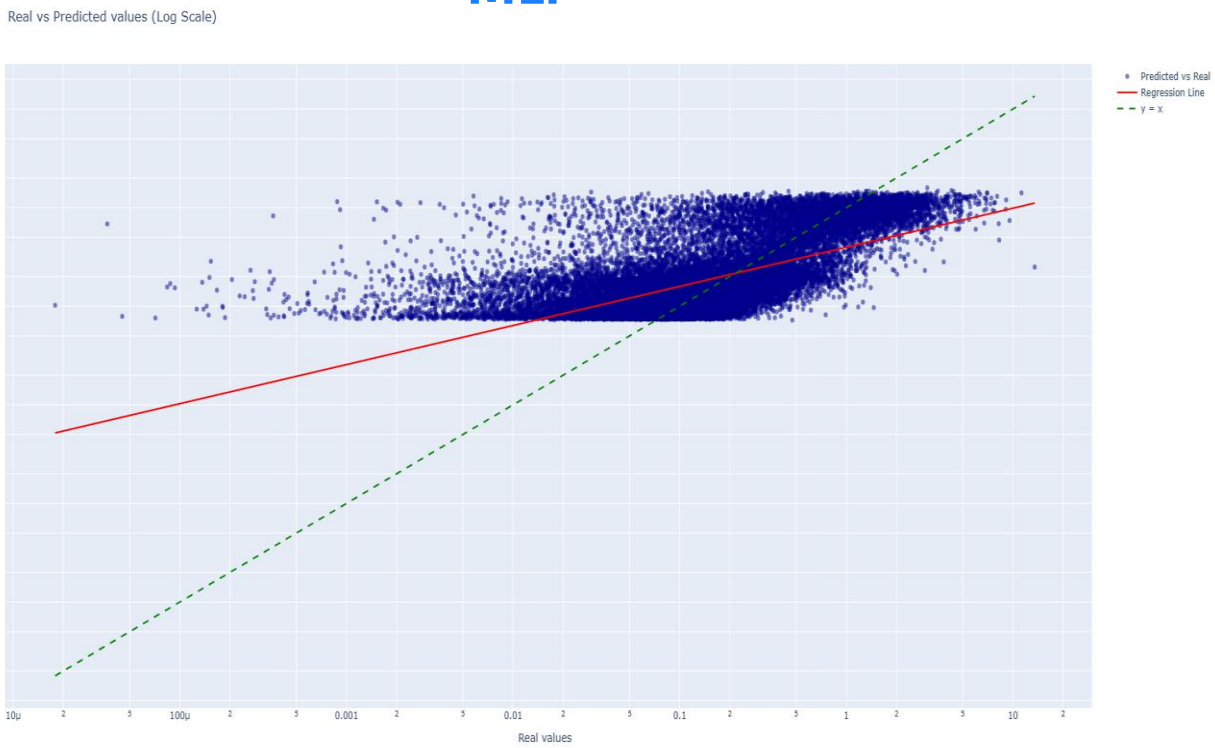


Mean Squared Error (MSE): 3.176 x 10 $^{-2}$

Mean Absolute Error (MAE): 9.104 x 10 $^{-2}$

Coefficient of determination (R^2): 2.254 x 10 $^{-1}$

MLP



Mean Squared Error (MSE): 3.075 x 10 $^{-2}$

Mean Absolute Error (MAE): 9.233 x 10 $^{-2}$

Coefficient of determination (R^2): 2.5 x 10 $^{-1}$

Removing the parallax error related fields caused a significant drop in performance: the previous results were heavily influenced by these fields.

Data Preparation – Separate datasets

Analyze the Cepheid and RRLyrae datasets separately and extract more fields with better correlations.

Because of the sparsity of the datasets more attention was given to preprocessing the data, removing outliers and rows with few information fields.

Name	Description
zp_mag_g	Zero point (mag) of the final model of the G-band light curve for Cepheids and RR Lyrae stars.
zp_mag_bp	Zero point (mag) of the final model of the GBP band light curve for Cepheids and RR Lyrae stars.
peak_to_peak_g	This parameter is filled with the peak-to-peak amplitude value of the G band light curve. The peak-to-peak amplitude is calculated as the (maximum) - (minimum) of the modelled folded light curve in the G band.
peak_to_peak_bp	Peak-to-peak amplitude of the BP band light curve (float, Magnitude[mag])
fund_freq1	First frequency of the non-linear Fourier modelling. It applies to all three G, GBP, and GRP bands and the radial velocity curve.
...	

Cepheid used dataset fields

Name	Description
astrometric_gof_al	Goodness-of-fit statistic of the astrometric solution for the source in the along-scan direction. This is the 'gaussianized chi-square', which for good fits should approximately follow a normal distribution with zero mean value and unit standard deviation. Values exceeding, say, +3 thus indicate a bad fit to the data.
astrometric_chi2_al	Astrometric goodness-of-fit (χ2) in the AL direction. χ2 values were computed for the 'good' AL observations of the source, without taking into account the astrometric excess noise (if any) of the source. They do however take into account the attitude excess noise (if any) of each observation.
astrometric_excess_noise	This is the excess noise of the source. It measures the disagreement, expressed as an angle, between the observations of a source and the best-fitting standard astrometric model (using five astrometric parameters).
...	

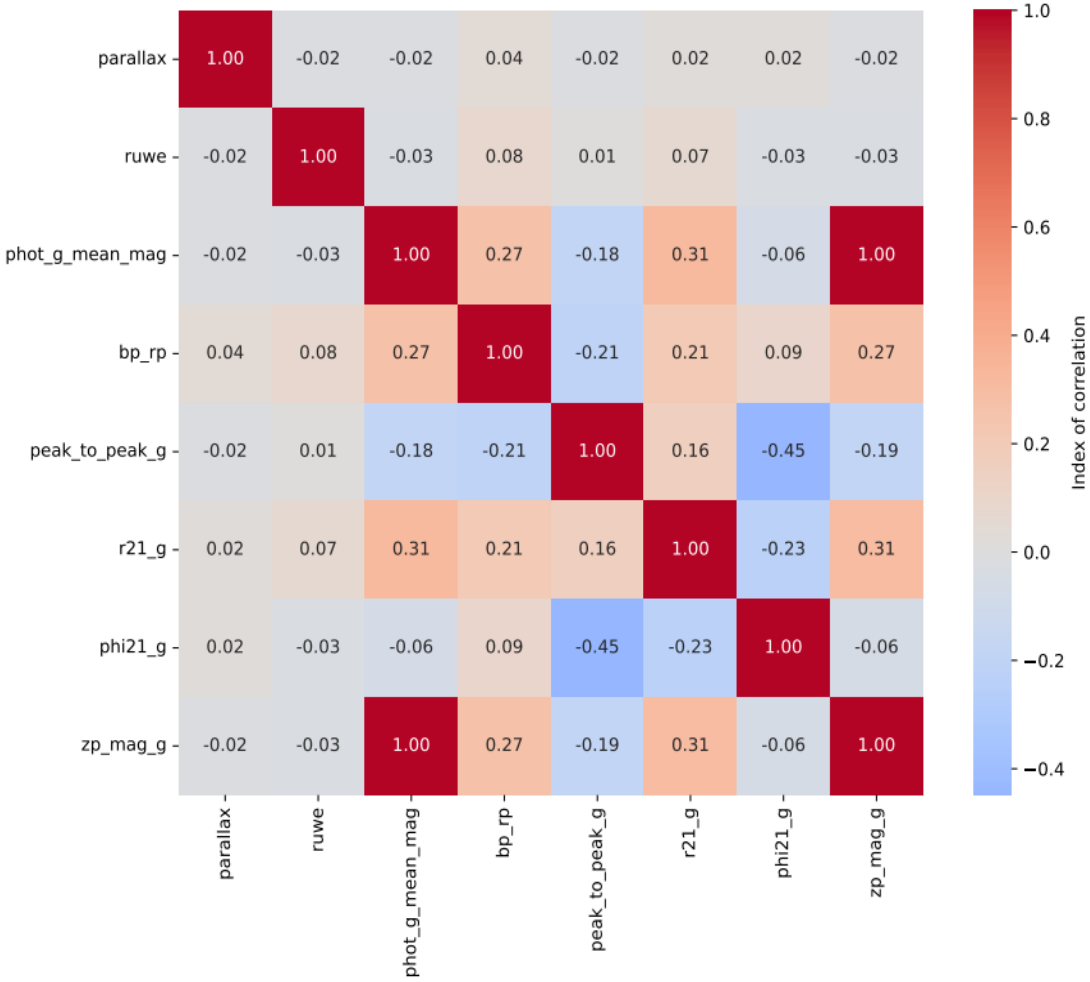
RRLyrae used dataset fields

Data Preparation – variables correlation (in progress)

Two variables highly correlated contain redundant information.
Keeping both can:

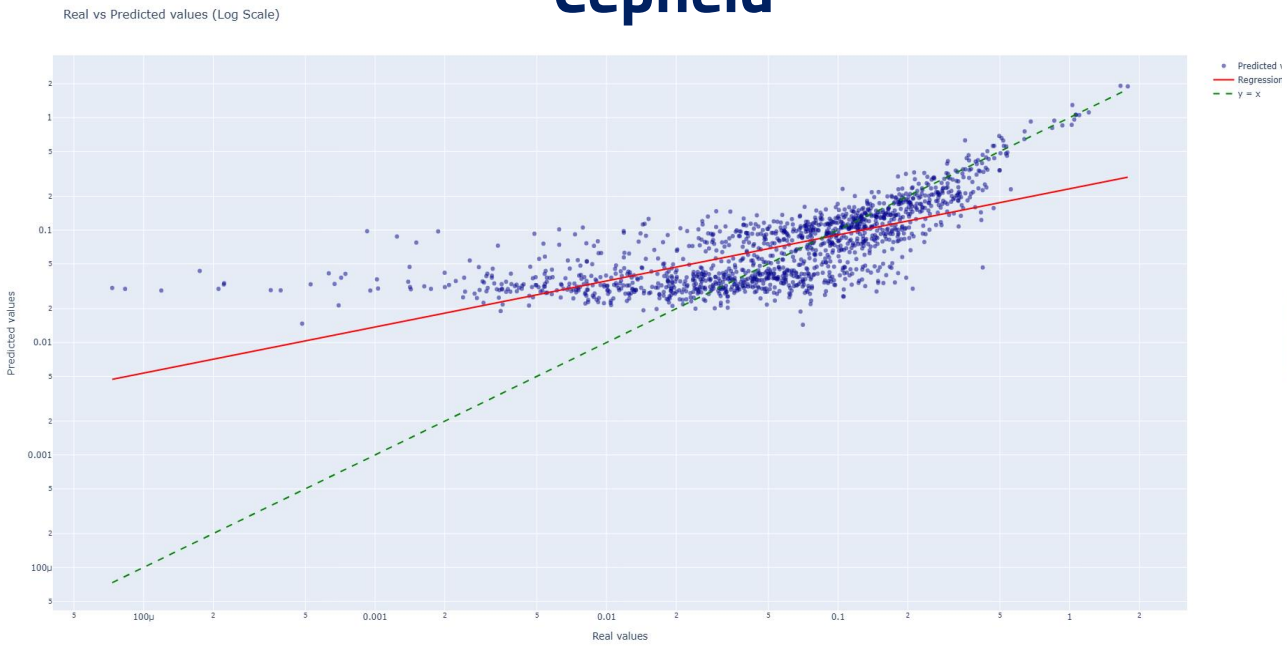
- Increase noise;
- Make the model more complex than necessary;
- Slow down training;
- Confuse the optimization process (weights "share" the effect);
- Worsen generalization.

- Variables highly correlated with the target: potentially useful;
- Variables weakly correlated with everything: may not carry useful information;
- Variables too strongly correlated with each other: useless, select only one.



Actual results (in progress)

Cepheid

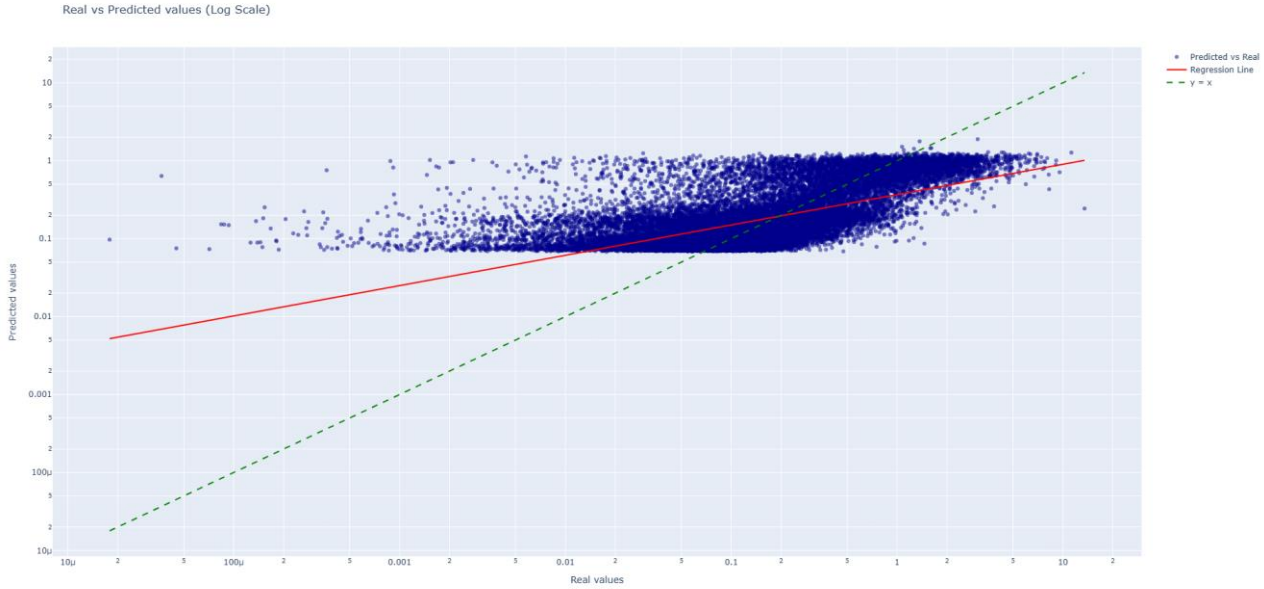


Mean Squared Error (MSE): 3.16505 x 10^{-3}

Mean Absolute Error (MAE): 3.86555 x 10^{-2}

Coefficient of determination (R^2): 8.46768 x 10^{-1}

RRLyrae



Mean Squared Error (MSE): 1.77831 x 10^{-1}

Mean Absolute Error (MAE): 3.68798 x 10^{-1}

Coefficient of determination (R^2): 5.22587 x 10^{-1}

After dataset separation and variables correlation study, results are slightly better than the previous ones.

Future development phases

Once the phases in progress will be completed:

Analysis and preprocessing of the dataset related to binary eclipses

Creation of a neural network for the parallax prediction of the binary eclipses

Creation of a neural network for the parallax prediction of the binary eclipses

Study of error propagation and management of uncertainties related to measurements for cepheid, RRLyrae and binary eclipses networks

Creation of a decision mechanisms based on simple averages, weighted averages or other techniques to be defined for cepheid, RRLyrae and binary eclipses networks

Creation of trained networks which can be used alone or in combination


Thank you!