ExoAID : Artificial Intelligence for exoplanet detection in the Gaia era

Asier Abreu [1,2] Jorge Lillo-Box [3] Ana-Maria Perez Garcia[3]
[1] ATG Europe for European Space Agency , ESAC 28692, Villanueva de la Cañada (Madrid), Spain. Asier.Abreu@ext.esa.int
[2] Universidad Complutense de Madrid, Av. Complutense, s/n, Moncloa - Aravaca, 28040 (Madrid).
[3] Centro de Astrobiología (CAB, CSIC-INTA), Depto. de Astrofísica, ESAC campus, 28692, Villanueva de la Cañada (Madrid). aperez@cab.inta-csic.es,jlillo@cab.inta-csic.es

1 Introduction

The current census of exoplanets has been almost entirely built upon transit and radial velocity detection methods. While astrometric exoplanet detection has been almost anecdotal so far due to the ultra-high (μas) precision required for this detection technique, the arrival of ESA's Gaia mission has radically changed the picture, paving the road for astrometric exoplanet detection at large and over the full-sky. Our project (ExoAID) aims to improve the current statistics and sky-coverage on detected exoplanets by application of machine learning and Artifical intelligence techniques to the Gaia catalog.

2 Methodology

We simulate the epoch astrometry (α_i, δ_i) $t_i \in [t_1, .., t_n]$ of a large (N=100000) set of stars using a single-star astrometric model[1] for the expected position of the star at time t:

$$\begin{aligned} \alpha^*(t) &= \alpha_0^* + \mu_{\alpha^*}(t - t_0) + \Pi_\alpha \varpi \\ \delta(t) &= \delta_0 + \mu_\delta(t - t_0) + \Pi_\delta \varpi \end{aligned} \quad (1)$$

Using Gaia scanning law we also simulate the measurements that Gaia would perform on our N stars at each t.

$$\eta_i(t) = \alpha_i^*(t) \sin\phi_t + \delta_i(t) \cos\phi_t \quad (2)$$

To a fraction of these stars we add one keplerian companion to create synthetic binary systems with different orbital characteristics $(a, e, i, P, t_p, \Omega, \omega)$. The presence of an unseen companion will shift the barycenter position and thus bias the actual observable (η_i^{obsv} = position of the binary photocenter at time t_i). This in turn will affect the quality of the least-squares fit[2] to the simulated astrometric observations:

$$\chi^2_{\text{red}} = \frac{1}{\nu} \cdot \sum_i \left(\frac{\eta_i^{\text{modl}} - \eta_i^{\text{obsv}}}{\sigma_i} \right) \quad (3)$$

For each star in this synthetic dataset we compute a 5-p astrometric solution $(\alpha^*, \delta, \mu_\alpha^*, \mu_\delta, \varpi)$ along with the corresponding astrometric quality fit statistics. We use the astrometric quality fit statistics as features to create a *labelled* dataset $\mathcal{D} = \{(\vec{x}_1, y_1), \ldots, (\vec{x}_k, y_k)\}$ composed of vectors of features $\vec{x}_i$ and their corresponding binary target y_i ($y = 0$ single star, $y = 1$ star with companion). We can therefore perform supervised learning using different type of ML classifiers to learn the underlying relation between features $\bar{x}_i$ and target y_i.

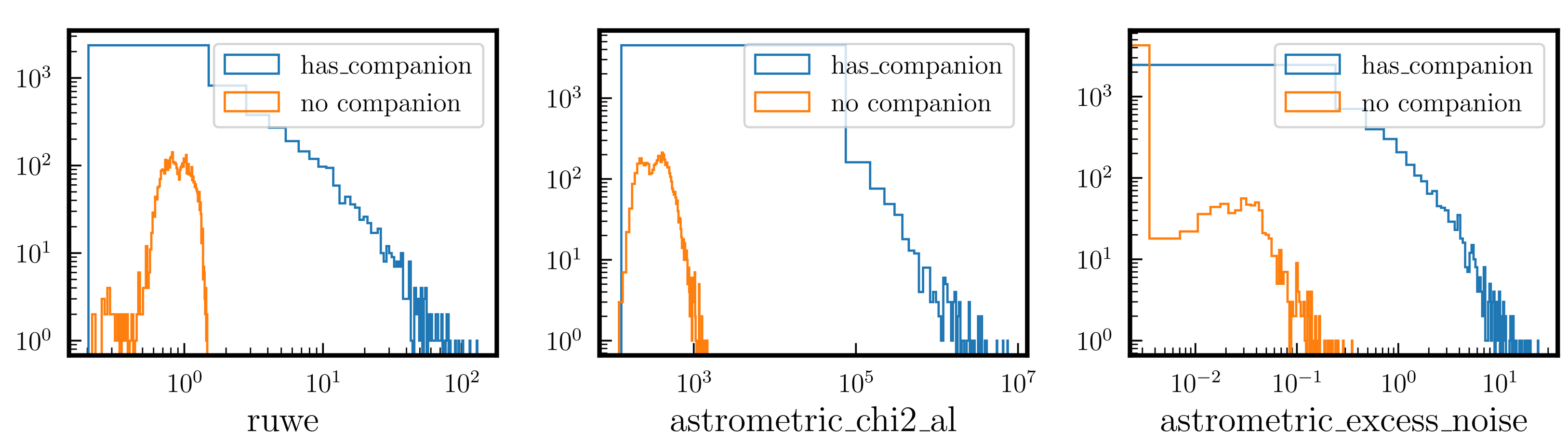

Figure 1: Distributions of fit quality parameters coming from the astrometric solution for the simulated systems, color coded by type of system (single=ORANGE or binary=BLUE), where the effect of the presence of a companion is clearly noticeable.

[1] where (α_0^*, δ_0) correspond to a reference position in the tangent observation plane at epoch t_0, $\alpha^* = \alpha\cos\delta$, ϕ_t is position angel of the scan at time t, Π_α,Π_δ are the source AL parallax factor components $(\partial\eta_\alpha/\partial\varpi, \partial\eta_\delta/\partial\varpi)$, and ϖ is the source parallax.

[2] the reduced chi-square metric is used as a goodness-of-fit metric, where $\nu = N_{observations} - 5$(degrees of freedom) and σ_i is the i-th measurement error.

3 Deep Learning

We choose to perform supervised learning using a Deep Neural Network consisting of 3 densely connected layers with 64 neurons on each layer.

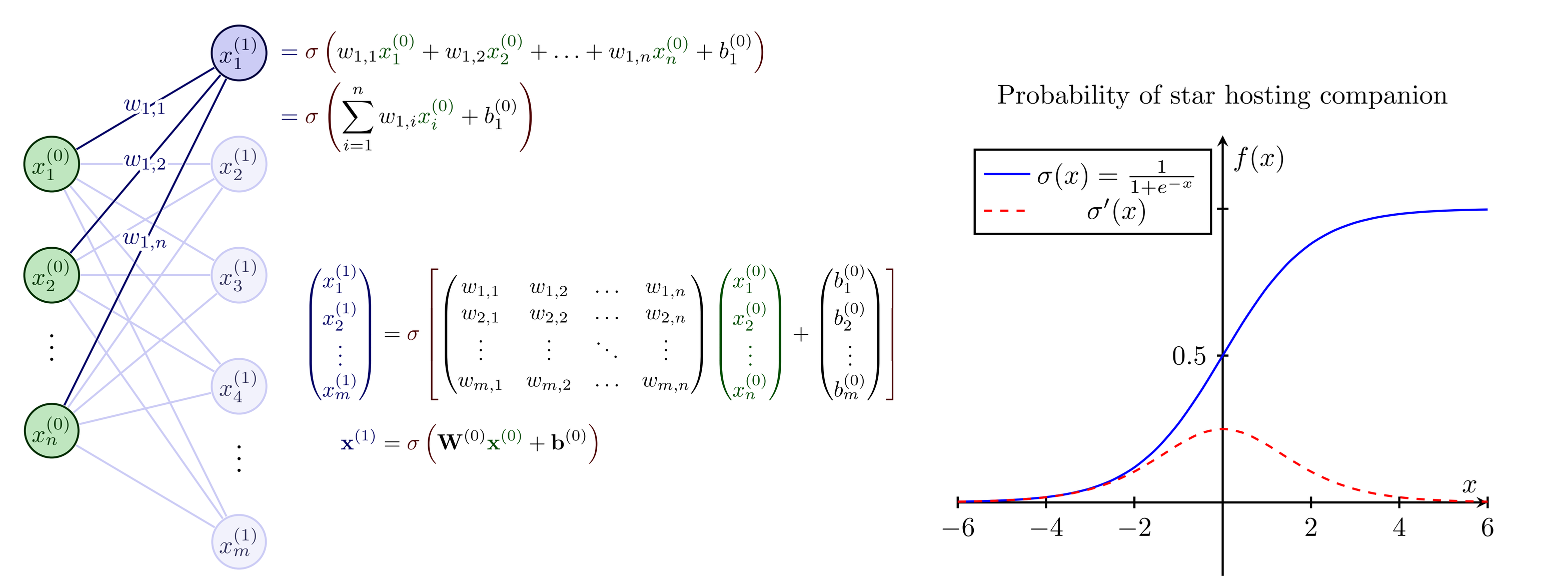

Figure 2: (LEFT) Schematic of the actual computation that takes place on ONE neuron of the first hidden layer our Deep Neural Network when an array of $\bar{x}$ features is given. The actvation function for the hidden neurons f is a ReLU.(RIGHT) The output layer of our DNN uses a sigmoid activation function to compute the probability of a star to host a companion.

We use ReLU activation on each hidden neuron, L2 regularization and a dropout layer to avoid over-fitting. We adopt a 70-30% train-test data split and train our DNN for 200 epochs. The loss function used during the DNN training is the binary cross-entropy.

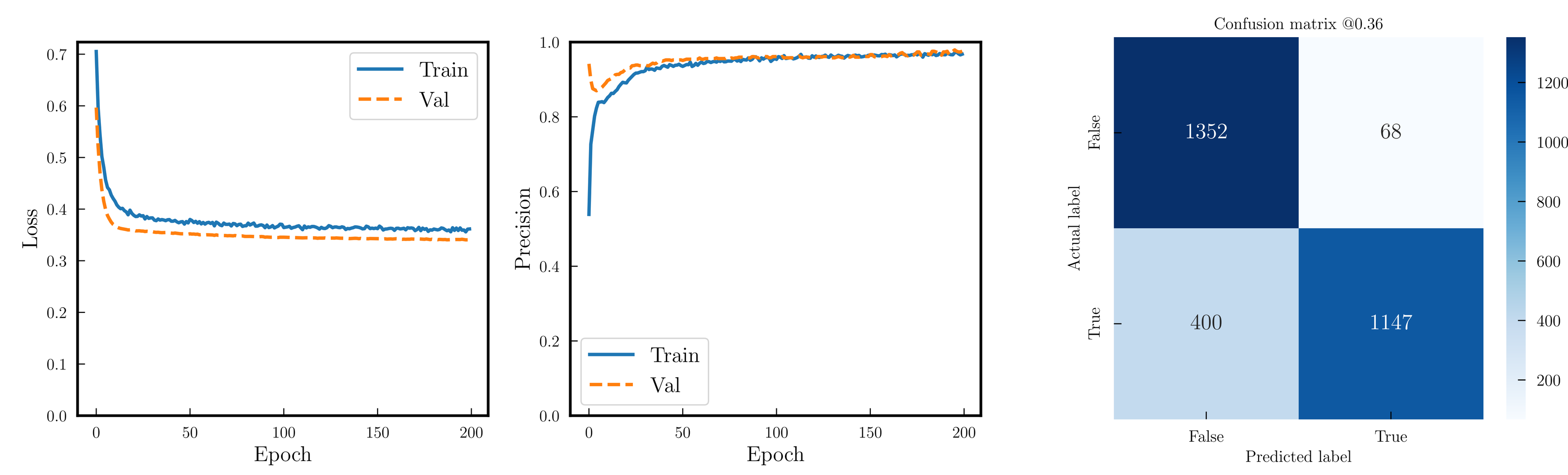

Figure 3: (LEFT) Training Loss and precision evolution during the training showing that the network is learning. (RIGHT) Confusion matrix corresponding to the test dataset showing the True Positives and Negatives along with False Positive and Negatives. The False Negatives (matrix bottom-left) are higher than the False Positives (matrix top-right), meaning our DNN predictions are not over-confident.

4 Model Predictions

We can apply the trained deep learning model to existing sources from the Gaia DR3 catalog and produce predictions of the probability of those stars to host a companion based on the values of the astrometric quality features as computed by the model.

Gaia Source Id	**Star Name**	$RUWE$	χ^2	ϵ	p
Gaia DR3 3026325426682637824	Gaia-1	1.052	662.486	0.087	0.877
Gaia DR3 1107980654748582144	Gaia-2	1.634	11564.604	0.199	0.997
Gaia DR3 2125960402948600704	Kepler-1534	1.652	786.902	0.550	1.000
Gaia DR3 2077382707923763328	Kepler-636	1.650	631.953	0.495	1.000
Gaia DR3 2134773160447382656	Kepler-365	1.641	670.894	0.544	1.000
Gaia DR3 5925209583053212800	GJ 676 A	1.641	2251.540	0.230	0.999
Gaia DR3 1107980654748582144	Gaia-2	1.634	11564.604	0.199	0.997
Gaia DR3 2785466581298775680	TOI-1468	1.625	1424.787	0.222	0.998
Gaia DR3 4050522981941484672	MOA-2007-BLG-400L	1.621	337.701	3.026	1.000
Gaia DR3 2135354424139022592	Kepler-1404	1.620	865.599	0.371	1.000
Gaia DR3 5065640460769428224	WASP-72	1.619	2819.416	0.189	0.996
Gaia DR3 4056155333341858304	OGLE-2012-BLG-0406L	1.612	939.444	1.284	1.000

Table 1: Example of model predictions for a small number of existing Gaia DR3 sources known to host exoplanets via other detection methods. Also included Gaia detected exoplanets via epoch astrometry. The columns correspond to the published astrometric quality fit statistics (RUWE,astrometric_chi2_al and astrometric_excess_noise) and the model computed probability to host a companion.

5 On-going Work

At present we are refining the DNN architecture and further calibrating the model against a set of reference datasets composed of stars belonging to the Non-Single-Stars Gaia catalog. Once this is completed we will perform large batch predictions on selected subsets of F,G,K,M stars to identify suitable candidates for follow-up Radial Velocity observations.