Photometric mass ratio of contact binaries with machine learning

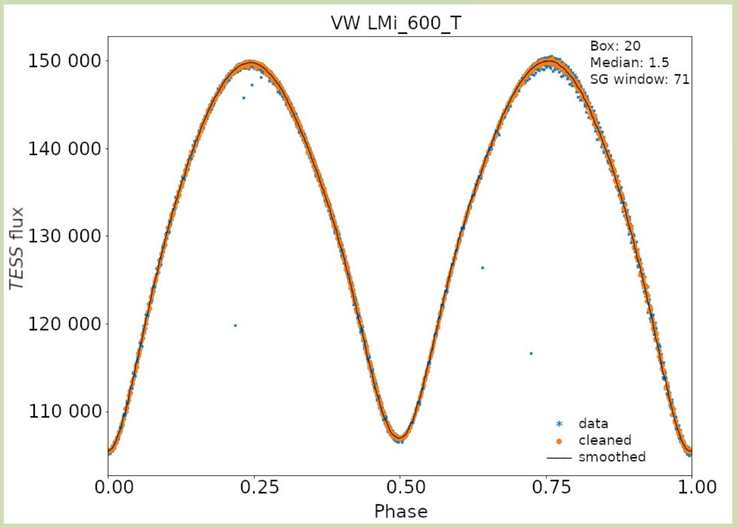
HAMBÁLEK, Ľubomír – MALIUK, Andrii – PRIBULLA, Theodor

Light curves (LCs) of close binary stars are relatively simple to describe when considering a small number of model parameters. However, systems with partial eclipses lead to a more complicated solution - because the mass ratio (q) of components correlates with the orbital inclination (i). In today's era of precise spaceborne photometry (e.g. *TESS*) we can spot subtle differences between individual light curves that could determine the investigated system's mass ratio (q), fill-out parameter (f), and inclination (i). We have created a database of synthetic light curves to train a simple machine learning model using Python's Sci-Kit to describe the shape of a normalized light curve, and then find probable values of q, i, and f parameters. We have also investigated the effect of unknown third light (l_3) in the system.

Preparation of data

We have selected 14 targets (see Tab. 2) with precisely measured mass ratio q_{spec} by spectroscopy such as to cover maximum range of q. Values of i, f, and l_3 were found in literature (for a complete list see [2]).

For our targets, we have found all available *TESS* data from all sectors with 120-s and 600-s cadence. From each individual sector, we have constructed a phased LC. We have used sigma clipping on a running box cart to eliminate outliers. Afterwards, we used a custom Savitzky-Golay filter (using SciPy) with parameters automatically adjusted for each LC to smooth the phased LC. The next step was to use our code `UNiQUE` [1] to normalize the LC and calculate a set of 11 (Fourier) coefficients a_i that represent the shape of the LC as:
$I(\varphi) = a_0 + \Sigma_i\, a_i \cos(2\pi i\varphi)$. All subsequent analysis was done according to these coefficients.



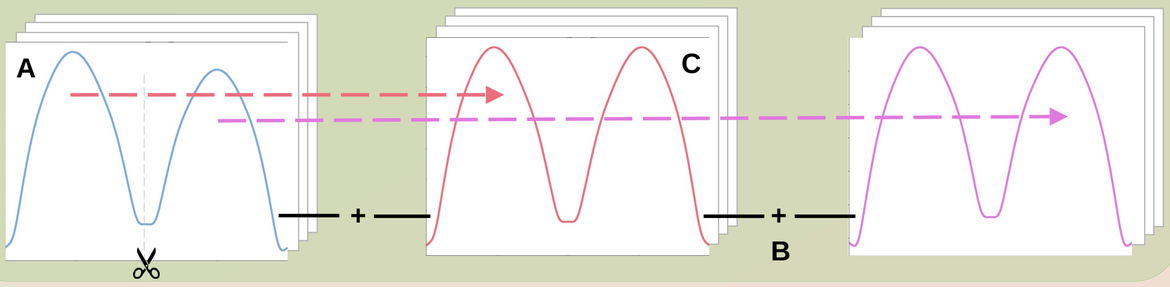
This work follows our previous attempt [2] to use only the code `UNiQUE` to find if we can sufficiently predict mass ratio q and inclination i of a close binary companions from the eclipsing LC.

Fig.1 – Illustration of the cleaning and smoothing process. *TESS* SPOC flux LC was cleared of outliers (blue) and the resulting LC (orange) was smoothed with Savitzky-Golay filter (with parameters in the upper right corner). This smoothed curve (black) was then fitted with the $I(\varphi)$ function to obtain the set of a_i coefficients.

Data sets

For all *TESS* LCs we have applied the procedure as described in the data preparation and ended up with Fourier coefficients a_i for each LC. Then we utilized our code `UNiQUE` (hereafter models U) in comparison to the `XGBoost` machine learning (hereafter models M). We have created several subsets (for both models):

- **A** - the initial set with normalized and detrended LCs,
- **B** - Because many LCs in set A were asymmetric in respect to the phase 0.5 (this was required for models U. Also models M were trained on symmetric data), we have extended set A by artificial LCs. LCs from set A were cut in half by phase φ = 0.5, and new LCs were created from phase parts $\langle 0, 0.5)$ and $\langle 0.5, 1)$ by mirroring around phase φ = 0.5.
- **C** - Since the asymmetry in LC maxima is caused by the O'Connell effect due to spots, for this subset, we have selected (from dataset B) artificial symmetric LCs with higher fluxes in maxima (as if the cold spot was not present).



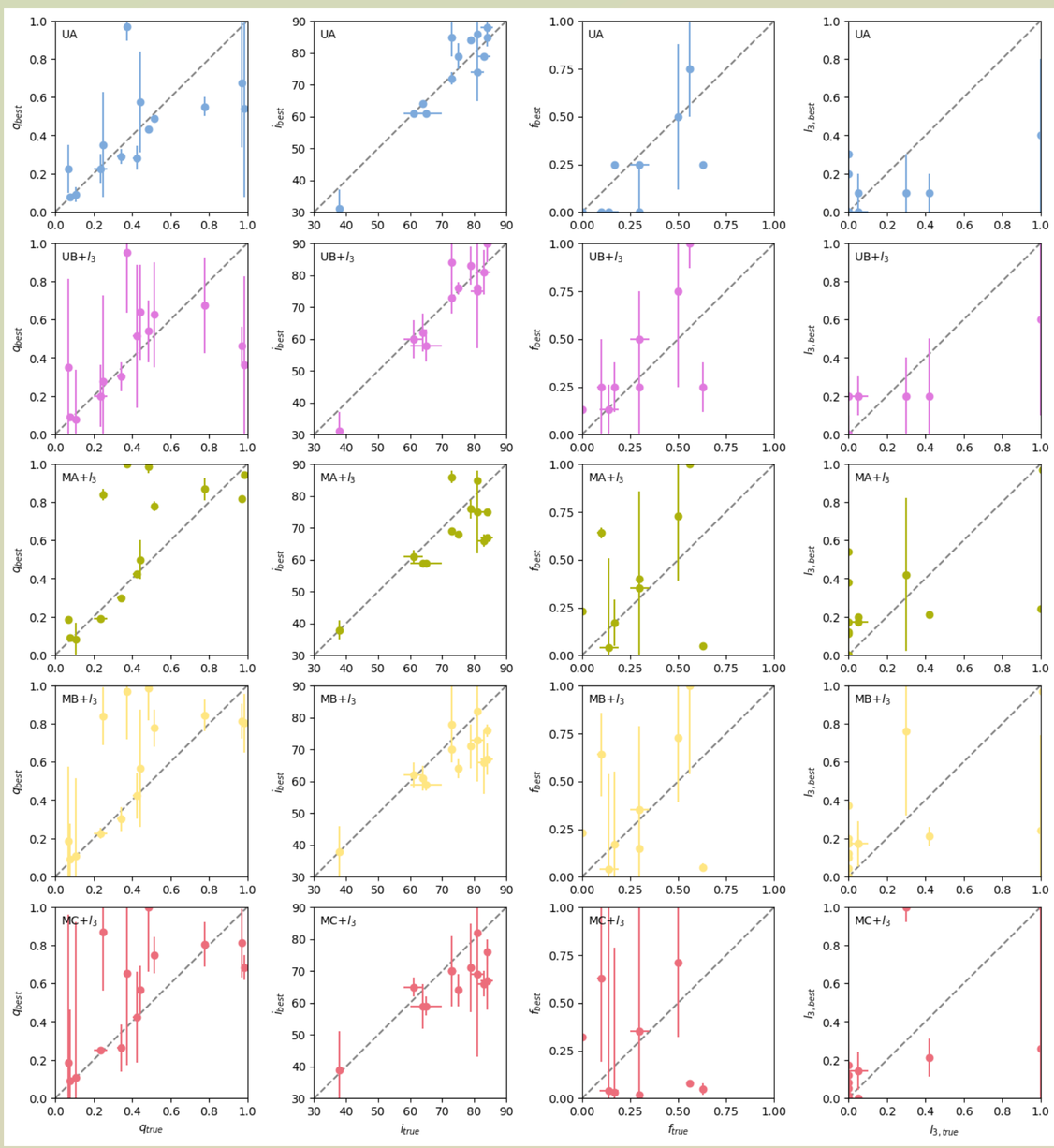


Fig. 3 – comparison of predicted values (y-axis) vs actual values (x-axis) of parameters (columns from left to right): mass ratio q, inclination i, fill-out factor f, and third light l_3 for different models (rows denoted by color) for individual stars (points). Compare with trained model in Fig. 2.

Model and training

Our pre-computed library for training consisted of 11,895 LCs saved as 11-tuple coupled with 3-tuple of parameters q, f, i. We have extended this set by another 5 levels of l_3 making it 71,370 entries in total. For better handling the parameter limits, we have converted the inclination i into $\sin(i)$. We chose the `XGBoost` model for its high performance, ability to handle heterogeneous tabular data effectively, and resistance to overfitting.

`XGBoost` builds a sequence of decision trees in a gradient boosting framework, where each tree corrects the errors of the previous one, rather than combining predictions from independent models. This approach allows `XGBoost` to capture complex nonlinear relationships between the input data and target variables, making it an optimal choice for regression tasks in astronomical research. The model was trained using a 70:30 train/test split, with the maximum tree depth $\mathrm{max\_depth} = 7$. This value was chosen by trial and error as a balance between capturing complex patterns in the data and avoiding overfitting.

To evaluate the model's performance, we used the root mean square error (RMS) as a metric. The results indicated successful training, with RMS values on the test set as follows: $q_{\mathrm{RMS}} = 0.01507$, $f_{\mathrm{RMS}} = 0.04081$, $\sin(i)_{\mathrm{RMS}} = 0.00566$, and third light contribution $l_{3,\mathrm{RMS}} = 0.05621$. For the training set, the RMS values were slightly better: $q_{\mathrm{RMS}} = 0.01084$, $f_{\mathrm{RMS}} = 0.02451$, $\sin(i)_{\mathrm{RMS}} = 0.00370$, and $l_{3,\mathrm{RMS}} = 0.03875$. The close match between training and test RMS values suggests that the chosen model complexity, guided by the $\mathrm{max\_depth}$ parameter, effectively prevented overfitting while ensuring high predictive accuracy (see Fig. 2).

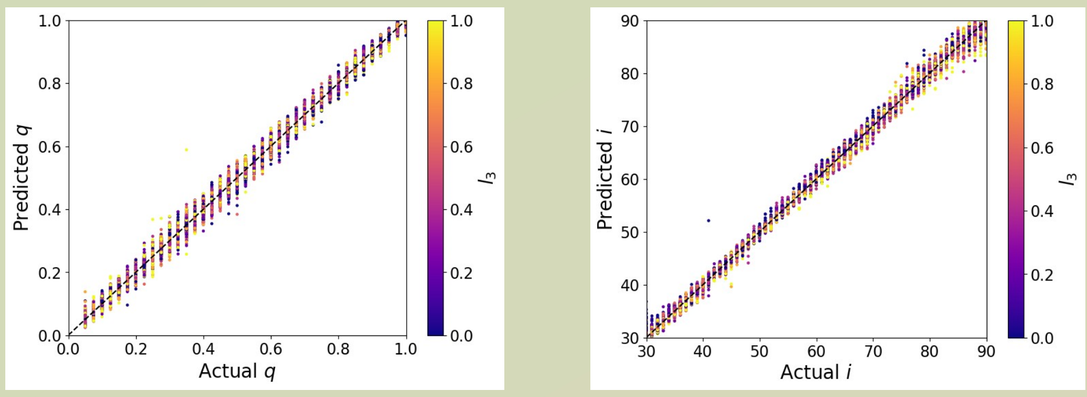


Fig. 2 – Testing the prediction of our `XGBoost` model for mass ratio q (left panel) and inclination i (right panel). On the x-axis we have values corresponding to all entries in the test subset, on the y-axis there are corresponding values predicted by the trained model. Colorbar codes different values of third light (l_3) contribution in the respective LC.

Results

We have compared all the parameters q, f, i, l_3 with those known in literature (see [1]) for all models. To compare, which model worked best for our sample of 14 stars, we have compared predicted "best" values and the spread of predictions (errorbars in Fig. 3), and computed correlation coefficients (see Tab. 1). If we are interested only in the mass ratio q, we have tabulated the predictions for individual targets (Tab. 2). The colors in tables and Fig. 3 are consistent with model used.

model	q	i	f	l_3
UA	0.978	0.955	0.708	0.083
UB+l_3	0.857	0.973	0.573	0.547
MA+l_3	0.787	0.952	-0.548	0.096
MB+l_3	0.839	0.896	-0.618	0.759
MC+l_3	0.897	0.853	-0.702	0.999

Tab. 1 – weighted corelation coefficients of predicted different parameters vs. "true" parameters.

Out of 14 stars in our sample, 6 systems are totally eclipsing. Only FT UMa and V753 Mon are classified as EB subtype. Our results show, that for systems with total eclipses, we get better predictions. So far we have not identified the single best approach in choosing a reliable model. Further analysis of O'Connell effect, shapes of minima of LCs, and larger sample of stars may be needed.

Object	q_{true}	UA	UB+l_3	MA+l_3	MB+l_3	MC+l_3
AG Vir	0.341(21)	0.288[38]	**0.300[75]**	0.296[8]	**0.300[62]**	0.261[123]
AW UMa	0.075(5)	**0.075[0]**	0.088[13]	0.089[10]	0.089[187]	0.089[374]
DU Boo	0.234(35)	**0.225[75]**	0.200[163]	0.190[0]	**0.226[30]**	**0.250[0]**
EL Boo	0.248(7)	0.350[275]	**0.275[450]**	0.839[30]	0.839[153]	0.869[306]
EQ Tau	0.442(10)	0.575[263]	0.638[250]	**0.498[103]**	0.566[308]	0.566[126]
FI Boo	0.327(9)	0.970[75]	0.950[313]	1.000[5]	0.970[252]	0.651[481]
FT UMa	0.984(19)	0.538[463]	0.363[463]	**0.941[14]**	0.802[153]	0.682[65]
SW Lac	0.776(14)	0.550[50]	0.675[250]	0.867[57]	0.843[82]	**0.804[115]**
SX Crv	0.066(3)	0.225[125]	0.350[463]	0.186[0]	0.186[388]	0.186[776]
V1191 Cyg	0.107(5)	0.089[38]	0.075[263]	0.083[85]	**0.105[408]**	**0.105[814]**
V523 Cas	0.516(8)	**0.488[13]**	0.625[275]	0.777[25]	0.777[98]	0.747[94]
V753 Mon	0.970(11)	0.675[338]	0.463[100]	0.817[9]	0.812[90]	0.812[180]
VW LMi	0.423(21)	0.281[63]	0.513[375]	**0.422[0]**	**0.422[119]**	**0.422[238]**
W UMa	0.484(3)	**0.433[13]**	0.538[163]	0.985[32]	0.985[170]	1.000[339]

Tab. 2 – medians of best predictions of mass ratio q in different models (colored background) for each of our test system with [(max - min)/2] value. **Bold**-face values indicate good match. Colors of models correspond to those in Fig. 3

Further reading:

[1] Hambálek, Ľ. & Pribulla, T., 2013: *CAOSP* **43**, 27
[2] Hambálek, Ľ. & Pribulla, T., 2024: *CAOSP* **54**, 175

Ľubomír Hambálek

lhambalek@ta3.sk
Astronomical Institute
Slovak Academy
of Sciences

This work was supported by:
APVV-20-0148
VEGA 2/0031/22