

Eclipsing binary classification with machine learning techniques

B. Keskin¹ and Ö. Baştürk^{2,3}

¹ *Ankara University, Graduate School of Natural & Applied Sciences,
Astronomy & Space Sciences Department, Ankara, Türkiye*

² *Ankara University, Faculty of Science, Astronomy & Space Sciences
Department, Ankara, Türkiye*

³ *Ankara University, Astronomy & Space Sciences Research & Application
Center, Kreiken Observatory, Ankara, Türkiye*

Received: November 14, 2024; Accepted: December 20, 2024

Abstract. We focus on the automated classification of eclipsing binary stars using deep learning methods to handle the vast data generated by large-scale photometric sky surveys. These surveys produce extensive datasets that are impractical for manual analysis. By using machine learning to classify eclipsing binary stars based on light curve morphology, this study aims to contribute to the efforts to efficiently process and accurately interpret massive data from projects Kepler, TESS and Gaia missions.

Key words: stars:binary stars, stars:eclipsing binaries, techniques:light curve classification, techniques:machine learning

1. Introduction

Eclipsing binaries, whose light curves show brightness variations from mutual eclipses, have large sets of photometric survey data. Automated data processing, leveraging supervised and unsupervised machine learning (ML) are essential to efficiently analyze these massive datasets and identifying patterns in time-series data.

Daza-Perilla et al. (2023) utilized ML to classify eclipsing binary stars (EBs) in the VISTA Variables of the Vía Láctea Survey (VVV), revealing time-series features in light curves and introducing a Compound Decision Tree (CDT) model for their classification. Čokina et al. (2021) classified eclipsing binary light curves into detached and over-contact classes. A hybrid of Bidirectional Long Short Term Memory (BiLSTM) and one dimensional Convolutional Neural Network (1D CNN), achieved 98% accuracy, and reached 100% when semi-detached binaries were excluded. Bódi & Hajdu (2021) used the Locally Linear Embedding (LLE) algorithm for classifying the Optical Gravitational Lensing Experiment (OGLE) eclipsing binary light curves based on their morphology. Süveges, M. et al. (2017) applied ML methods, including Functional Principal Component

Analysis (FPCA), Linear Discriminant Analysis (LDA), Random Forest (RF), and Self-Organizing Map (SOM), to classify eclipsing binaries based on light curve morphology using datasets from Catalog and Atlas of Eclipsing Binaries (CALEB), High Precision Parallax Collecting Satellite (HIPPARCOS), and Kepler. Kochoska, A. et al. (2017) proposed a combination of the t-distributed Stochastic Neighbor Embedding (t-SNE) and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithms for the purposes of eclipsing binary light curve classification. The polynomial chain (polyfit) and two-Gaussian models are used to characterize the geometry of the folded light curves. Classification is done according to the morphology parameter for a given system.

This study aims to contribute to the efforts in reliable and rapid classification of eclipsing binaries, enabling statistically reproducible results across vast astronomical databases. We classify a limited sample of eclipsing binary star candidates in the Gaia DR3 archive. We utilize Transiting Exoplanet Survey Satellite (TESS)¹ and Kepler² light curve data in the Villanova eclipsing binary catalogs, where the systems are ready to be labeled according to the morphology parameter (Matijević et al., 2012), to train our CNN model. We then apply the trained model to classify our sample of eclipsing binaries from the Gaia Data Release 3 (DR3) archive.

2. Data and methods

Villanova Kepler and TESS eclipsing binary archives are used for training ML model. 2907 Kepler and 4349 TESS EB light curves are taken and labeled according to morphology parameter by the following criteria (Matijević et al., 2012);

- morph < 0.5 detached (D)
- 0.5 < morph < 0.7 semidetached (SD)
- 0.7 < morph < 0.8 overcontact (OC)
- 0.8 < morph ellipsoidal (E)

The light curves are phase-folded by using light elements taken from Villanova archives and binned uniformly to standardize their resolution, ensuring consistency across datasets (Fig.1).

For classification, a small sample including 2106 EB's were selected from Gaia DR3 archive. We provide a few of these binned and phase-folded light curves in Fig.2 (left) as examples. Light curves had to be processed based on two Gaussian and a cosine function due to their sampling and intensity scaling as presented in Mowlavi, N. et al. (2023) (Fig.2 right).

¹<https://tessebs.villanova.edu>

²<https://keplerebs.villanova.edu>

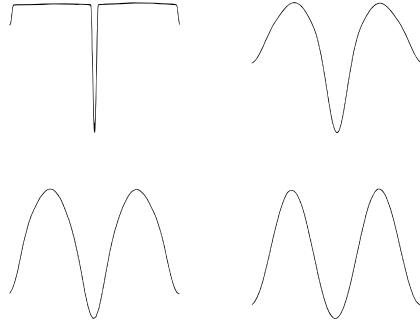


Figure 1. Sample detached (top left), semidetached (top right), overcontact (bottom left), ellipsoidal (bottom right) Kepler and TESS light curves.

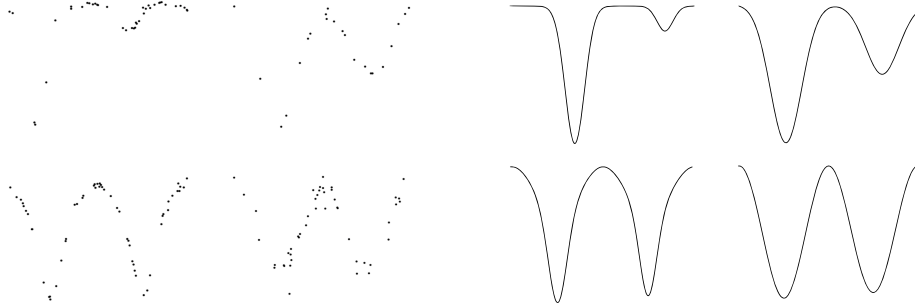


Figure 2. Sample Gaia DR3 light curves (left) and their modeled light curves (right).

The selected binaries are also in the Kepler or TESS archive. Out of 2106 EB's 789 are from Kepler archive and 1317 are from TESS archive. So we can crossmatch the labeled and the predicted classes.

We converted light curve data of Kepler and TESS EBs to Portable Network Graphic (PNG) image files by a Python code we developed for the task. Then these light curve images were splitted into 3 groups: 67% for training, 25% for validation and 8% for test. We used VGG-19 as the ML model. VGG-19 is a pretrained CNN from Visual Geometry Group (VGG) Department of Engineering Science, Oxford University. The number 19 stands for the number of layers with trainable weights. We chose VGG-19 for its ability to effectively extract nuanced features from light curve data, such as subtle variations in amplitude, shape and periodicity. VGG-19's fine-tuning, combined with the use of regularization techniques, provided results without overfitting. We used "reduce LR on plateau" and "early stopping" methods to avoid overfitting. We also conducted experiments with varying hyperparameters and confirmed that overtraining did not occur.

3. Results and discussion

We achieved 91% accuracy on Kepler and TESS test data and 64% accuracy on Gaia DR3 data as given with the confusion matrix in Fig.3. Our ML model is highly successful on semidetached class, while it tends to predict overcontact binaries as semidetached.

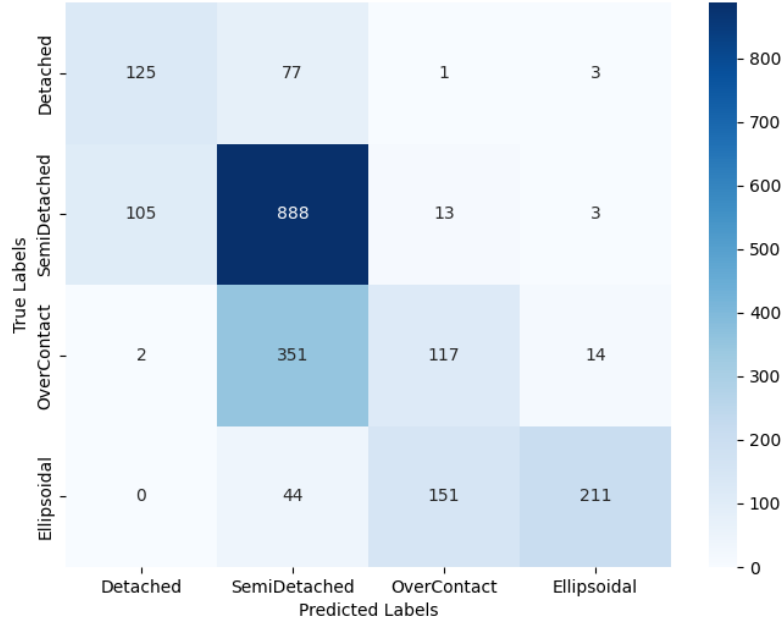


Figure 3. Confusion matrix.

Training of the ML model is one of the important steps. The Kepler and TESS light curves that we used for training should be modeled in the same way as the Gaia light curves. This is the reason why the model was successful in classifying the test data, but partially failed in classifying the Gaia data. We plan to explore additional modeling techniques, including smoothing and fitting analytic models, for improved consistency and interpretability in future studies.

Although model accuracy could probably be improved with a larger and more homogeneous dataset, our primary focus was to establish a proof of concept and identify potential challenges to this classification task. As part of our ongoing work, additional observation archives are being explored to expand the training dataset and improve the robustness and performance of the model.

We used "phase folding method" for the acquisition of the relevant light curve information. Harmonic analysis can be used to determine characteristics of the light curves. Integration of spot-induced effects as well as modulations

caused by reflection and ellipsoidal deformation in close binary systems could enhance the precision of future models, particularly for more detailed analysis of system parameters. Spot-induced variations are generally low-amplitude compared to the primary eclipse features in eclipsing binaries. Despite such variations, classification accuracy is found to be similar across diverse types of light curves with varying levels of spot activity.

References

- Bódi, A. & Hajdu, T., Classification of OGLE Eclipsing Binary Stars Based on Their Morphology Type with Locally Linear Embedding. 2021, *The Astrophysical Journal Supplement Series*, **255**, 1, DOI:10.3847/1538-4365/ac082c
- Čokina, M. and Maslej-Krešňáková, V., Butka, P., & Parimucha, Š., Automatic classification of eclipsing binary stars using deep learning methods. 2021, *Astronomy and Computing*, **36**, 100488, DOI:10.1016/j.ascom.2021.100488
- Daza-Perilla, I. V., Gramajo, L. V., Lares, M., et al., Automated classification of eclipsing binary systems in the VVV Survey. 2023, *Monthly Notices of the Royal Astronomical Society*, **520**, 828, DOI:10.1093/mnras/stad141
- Kochoska, A., Mowlavi, N., Prša, A., et al., Gaia eclipsing binary and multiple systems. A study of detectability and classification of eclipsing binaries with Gaia. 2017, *A&A*, **602**, A110, DOI:10.1051/0004-6361/201629957
- Matijević, G., Prša, A., Orosz, J. A., et al., Kepler Eclipsing Binary Stars. III. Classification of Kepler Eclipsing Binary Light Curves with Locally Linear Embedding. 2012, *The Astronomical Journal*, **143**, 123, DOI:10.1088/0004-6256/143/5/123
- Mowlavi, N., Holl, B., Lecoœur-Taïbi, I., et al., Gaia Data Release 3 - The first Gaia catalogue of eclipsing-binary candidates. 2023, *A&A*, **674**, A16, DOI:10.1051/0004-6361/202245330
- Süveges, M., Barblan, F., Lecoœur-Taïbi, I., et al., Gaia eclipsing binary and multiple systems. Supervised classification and self-organizing maps. 2017, *A&A*, **603**, A117, DOI:10.1051/0004-6361/201629710