SHAPE CHARACTERIZATION AND TEMPORAL ANALYSIS OF EXOPLANET TRANSITS WITH DEEP LEARNING.
C. J. Hönes[1,2,5], A. M. Heras[3], B. Foing[3,5,6,7], Y. Rusticus[2,5], V. Foing[2] and J. M. Terpstra[4,5]
[1] christoph.hoenes@gmail.com, [2] Universiteit van Amsterdam Department of Artificial Intelligence, [3] ESA ESTEC, [4] University College Amsterdam, [5] VU Amsterdam, [6] Leiden Observatory, [7] ILEWG EuroMoonMars

Introduction: Since the discovery of the first exoplanet more than 4,000 exoplanets have been found and many thousands of planetary candidates are yet to be confirmed. This rapid progress was possible thanks to several missions targeted on finding new exoplanets. In particular the *Kepler/K2* mission which alone discovered over 2,600 exoplanets [1]. The *TESS* mission [2] continues its legacy and is believed to discover many more.

The amount of collected data is already large and the rate at which new data becomes available is still increasing rapidly. This goes far beyond the capacities of manual inspection by human experts. The use of Artificial Intelligence (AI) methods, especially Deep Learning (DL), as an automated way of screening the data for potentially interesting planetary candidates, have lately gained great popularity and are already established as an essential tool for exoplanet detection.

These techniques are powerful means for pattern recognition and their applicability for exoplanetary science is not limited to the detection of exoplanetary candidates. This work presents an approach to automatically characterize transit shapes of potential exoplanetary signals and to perform a temporal analysis of the variations in the transit shapes over multiple consecutive transits of the same target. We hope to find anomalies in the transit shapes that can hint at interesting astrophysical phenomena such as disintegrating planets, gravity darkening or other asymmetries in the transit shapes. Additionally, we will try to draw conclusions about the underlying stellar activity of the host star by monitoring short- and long-term variability in the appearance of the transit shapes of the same target.

Data Preparation: For our experiments we are using light curve data collected by the TESS mission. We focus on the subset of TESS objects of interest (TOIs). As our research focuses on transit shapes, we crop the data to keep only the relevant (in-transit) part of the light curve based on the estimates of the epoch, orbital period and transit duration of the light curve provided by the TESS pipeline [3].

We consider different levels of pre-processing. The raw flux, the pipeline corrected flux and the pipeline corrected flux after detrending was applied to the whole light curve. The trade-off to be made here is between removing systematic noise, due to instrumental mechanisms, and preserving all relevant true signals that can enhance the quality of the analysis. Providing our approach with multiple granularities of pre-processing enables our method to dynamically make this decision by itself. Additionally, we incorporate time information to include knowledge about the temporal distance between subsequent transits and data points.

Method: A popular DL approach for the detection of exoplanets are variants of *Astronet* [4]. It is based on a one-dimensional convolutional neural network (CNN) which applies several filters to the data that are learned via the stochastic gradient descent (SGD) optimization algorithm. These filters extract information relevant to achieve the training objective, which was in that case distinguishing planetary transits from false positives.

For the characterization of transit shapes, we use a similar CNN-based architecture. However, our approach differs substantially from that of Astronet, since we do not seek to differentiate real transits from false positives, but rather to extract information from the data that can be used to characterize the shape of a transit. We design an autoencoder (AE) architecture

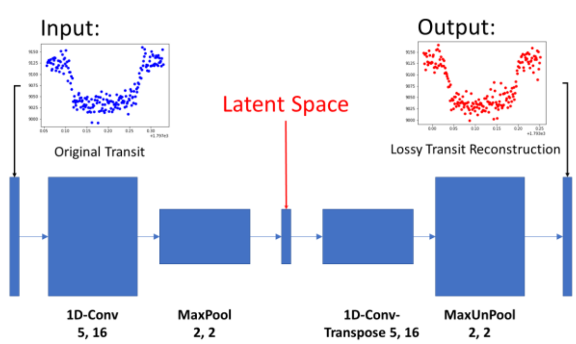

Figure 1: Sketch of the AE architecture. (Left) the encoder that encodes the shape of a transit into a latent space. (Right) the decoder that reconstructs the transit from the latent representation. The two numbers below the Conv layers represent the kernel size and number of filters respectively. For the Max(Un)Pool layer they represent the kernel size and the stride. Note that this is a simplified illustration which does not show all the convolutional blocks.

that is trained to first compress the data to a lower dimensional representation to subsequently reconstruct the original transit from this representation. This ensures that the model learns distinguishing features of the transit shape necessary to reconstruct the original data. An illustration of the shape model architecture can be seen in Figure 1.

The extracted feature representation will be used as input to our time series analysis model. For this we use *Long Short-Term Memory networks* (LSTMs). This method is suited for capturing trends in sequential data. It takes as input the data of the current time step as well as the output of the previous time step. An illustration of this architecture can be seen in Figure 2.

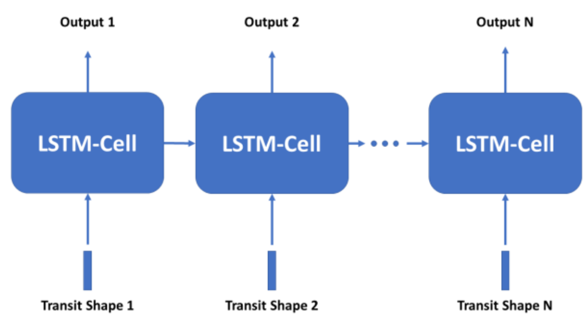

Figure 2: Sketch of an LSTM architecture that gets the representation of N transit shapes as input.

First Results and Discussion: The transit shape embedding learned by the shape model can be visualized by projecting the learned representations of the different transits to a two-dimensional manifold via dimensionality reduction methods like *t-distributed Stochastic Neighbor Embeddings* (t-SNE) [5]. The manifold created from a small subset of our data is shown in Figure 3.

The figure shows that similar transit shapes from the same target are clustered in close proximity to each other in the manifold. This indicates that the model has learned to capture features well-suited to describe transit shapes. With those features one can find similar examples and distinguish different types of transit shapes from each other.

By flagging certain known phenomena similar effects can be found for other targets, by retrieving the nearest neighbors in the embedding space. Additionally, new phenomena could be discovered by manually inspecting examples from clusters which are distant from flagged clusters and thereby cannot be attributed to any other known phenomena.

The further development and experiments for the time series analysis are a work in progress but we hope to be able to track the stellar variability over time using our LSTM architecture.

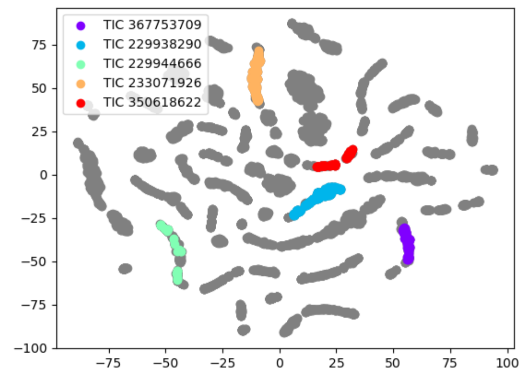

Figure 3: t-SNE manifold of transit shape embedding of 246 targets (6100 transits) where each data point represents one transit (in grey). For five randomly selected targets all transits are plotted in a certain color.

Future Work: A possible idea for enhancement is the incorporation of domain knowledge in the form of external stellar or planetary parameters such as surface temperature of the host star or the ratio of planetary and stellar radii. Additionally, the AE architecture could be extended to a Variational Autoencoder (VAE) [6] to ensure continuity of the latent space representation.

Acknowledgments: We want to thank Benjamin Miller for helpful discussions and his supervision on the Machine Learning part of this work. This research has made use of the NASA Exoplanet Archive, which is operated by the California Institute of Technology, under contract with the National Aeronautics and Space Administration under the Exoplanet Exploration Program.

References:
[1] NASA Exoplanet-Catalog (2021, January 07), from: https://exoplanets.nasa.gov/discovery/exoplanet-catalog/, [2] Ricker et al. (2014), Journal of Astronomical Telescopes, Instruments, and Systems (JATIS), 1.1.014003, [3] Jenkins J. M. et al. (2016), Proceedings of the *SPIE*, Volume 9913, id. 99133E 20 pp., [4] Shallue, C. J., & Vanderburg, A. 2018, AJ, 155, 94., [5] Maaten, L. v. d. & Hinton, G. (2008), Journal of Machine Learning Research, 9, 2579., [6] Kingma, D. P. & Welling, M. (2014), ICLR