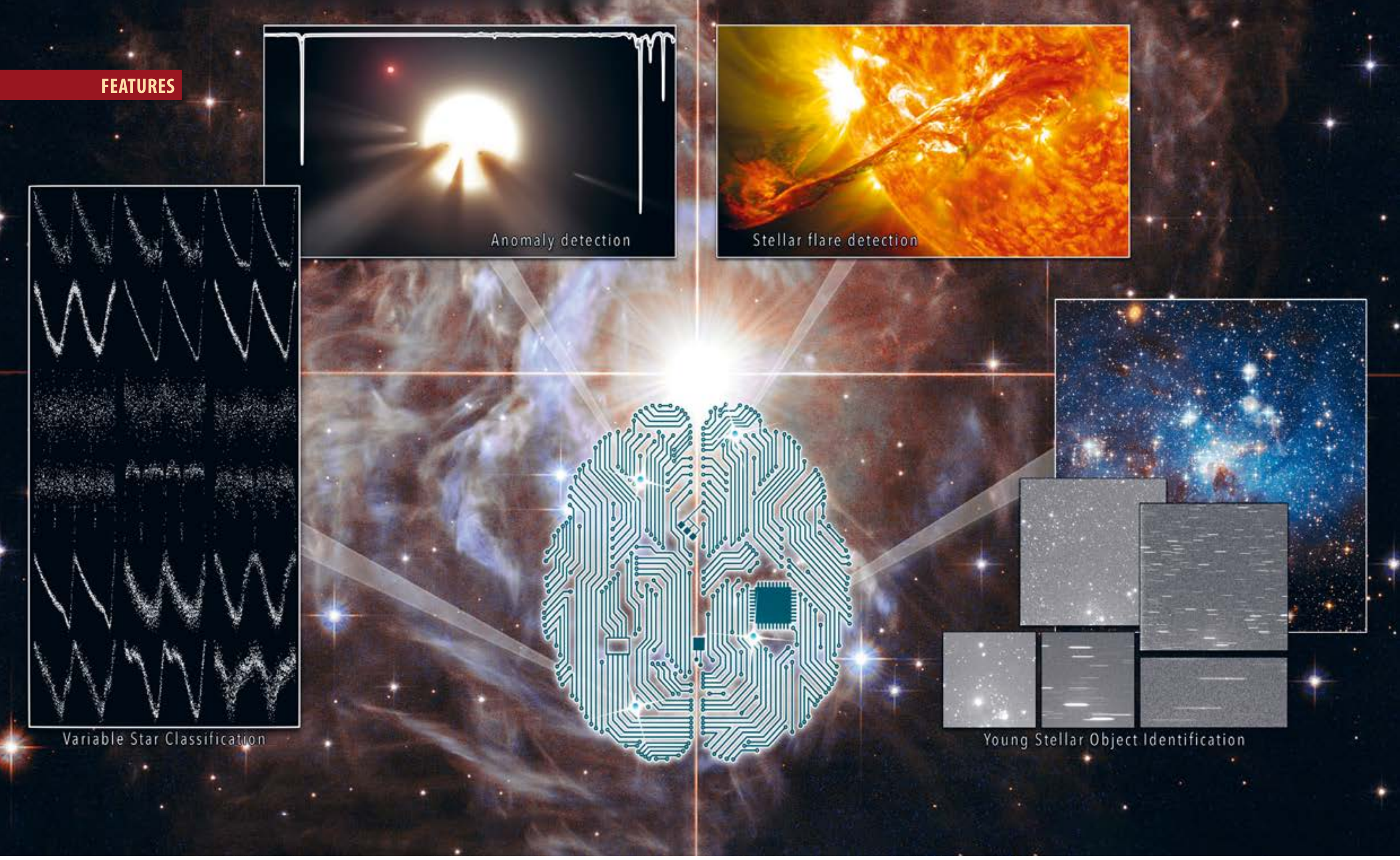


MACHINE LEARNING IN PRESENT DAY ASTROPHYSICS

R. Szabó[1,2,3], T. Szklenár[1,2], A. Bódi[1,2] – DOI: https://doi.org/10.1051/epn/2022205
[1] CSFK Konkoly Observatory, H-1121 Budapest, Konkoly Thege Miklós út 15-17., Hungary
[2] MTA CSFK Lendület Near-Field Cosmology Research Group
[3] ELTE Eötvös Loránd University, Institute of Physics, Budapest, Hungary

Machine learning is everywhere in our daily life. From the social media and bank sector to transportation and telecommunication, we cannot avoid using it, sometimes even without noticing that we are relying on it. Astronomy and astrophysics are no exception. From telescope time and survey telescope scheduling through object detection and classification, to cleaning images and making large simulations smarter and quicker to it is ubiquitous to use machine learning algorithms. To illustrate this silent revolution, we checked the NASA Astronomical Data System website[1] and searched for the keyword 'machine learning' in abstracts of astronomical and astrophysical papers. In 2000 we found 56, in 2010 889, and by 2020 no less than 35,659 abstracts contained the magic two words.

[1] https://ui.adsabs.harvard.edu/

▲ Typical applications of machine learning techniques in astronomy from object detection and classification to finding anomalies.

No wonder, since existing and upcoming astronomical databases are truly 'astronomical': Vera C. Rubin Observatory's Legacy Survey of Space and Time[2] will deliver 150 Petabytes of photometric data and images in the optical and near-infrared wavelength range, while the upcoming Square Kilometer Array[3] will produce 5 Zettabytes of data in the radio domain by 2030, just to mention two soon-to-be-online large surveys. To keep up with, exploit, and understand this tsunami of data, applying machine learning is a must. In this paper we highlight a few interesting cases in astronomy admitting that because of the breadth of this topic only a subjective selection is possible.

> **"Our research team has developed high-precision algorithms to classify the light curves of periodic variable stars into a few main classes."**

Anomaly detection, let's find the 'unknown unknowns'

The discovery of new astrophysical phenomena has long been the major goal of astronomical research. In astronomy a new physical phenomenon can manifest itself in many forms, from strange-looking shapes in images to exotic behaviours in time-series observations. With the advent of large-scale sky surveys such as Gaia, Pan-STARRS, ZTF, LSST, not only the well-known stars are being regularly observed, but everything that is bright enough to be detected with few-meter-class telescopes. Thanks to these observations, it has become possible to discover such intriguing bodies as 'Oumuamua, the first known interstellar object that passed through the Solar System, or the Boyajian star, the 'most mysterious star in the Universe', whose brightness variation is so unusual that at one point even alien megastructures were invoked to explain the observations. However, due to the exponentially growing amount of data, the traditional, human supervised way of data processing is less and less feasible, indicating the necessity of development of automatic novelty detection methods. The increasing pressure to enter the era of big data has led the astronomical community to utilise machine learning techniques to look for out-of-distribution anomalous astronomical objects.

A few recent examples for novelty or anomaly detection in astrophysics: [1] processed the raw data of the Open Supernova Catalog and identified non-supernova events and representatives of the rare supernova classes. [2] proposed and demonstrated a new anomaly detection technique to discover anomalous X-ray sources via high-resolution spectroscopy. [3] proposed a method to find a second Earth, *i.e.*, to detect potentially habitable exoplanets as anomalies. Finally, [4] used the light curve of periodic variable stars and identified stars with irregular variability. The one-million-dollar question is of course how to find the needle in the haystack, *i.e.*, how to find and follow-up(!) the rarest and/or most interesting (maybe Nobel-prize winning) objects among the million(s) of transients detected by LSST on a given night when the 8-meter large field-of-view Simonyi telescope starts surveying the Chilean sky in 2023.

Classification – let's teach the computer to 'see' like a human

Classification is the guinea pig of machine learning applications in astronomy. Large sky surveys conveyed millions or billions of images, spectra, positions, and proper motions of stars and galaxies and other celestial bodies. To make sense of these data one needs help from the 'silicon brains.' A subfield of astrophysics – but nonetheless a very important topic – is studying variable stars. These stars allow an unprecedented view into stellar interiors, help to establish the cosmic distance scale, and can be used as tracers of Galactic formation and evolution. Astronomers traditionally have classified objects that vary their brightness based on their light curves, that is graphical representation of light variation as a function of time. If additional information (*e.g.*, a spectrum) is available, then certain degeneracies can be broken. Classification can be performed based on several mathematical parameters of light curve data (range, scatter, number of zero crossings, skewness, *etc.*), but what if we want to mimic the human brain and feed images of light curves into the computer? Well, our group did exactly that [5].

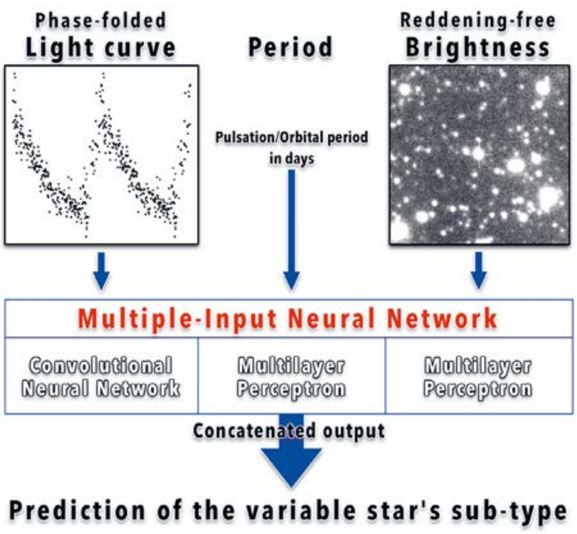
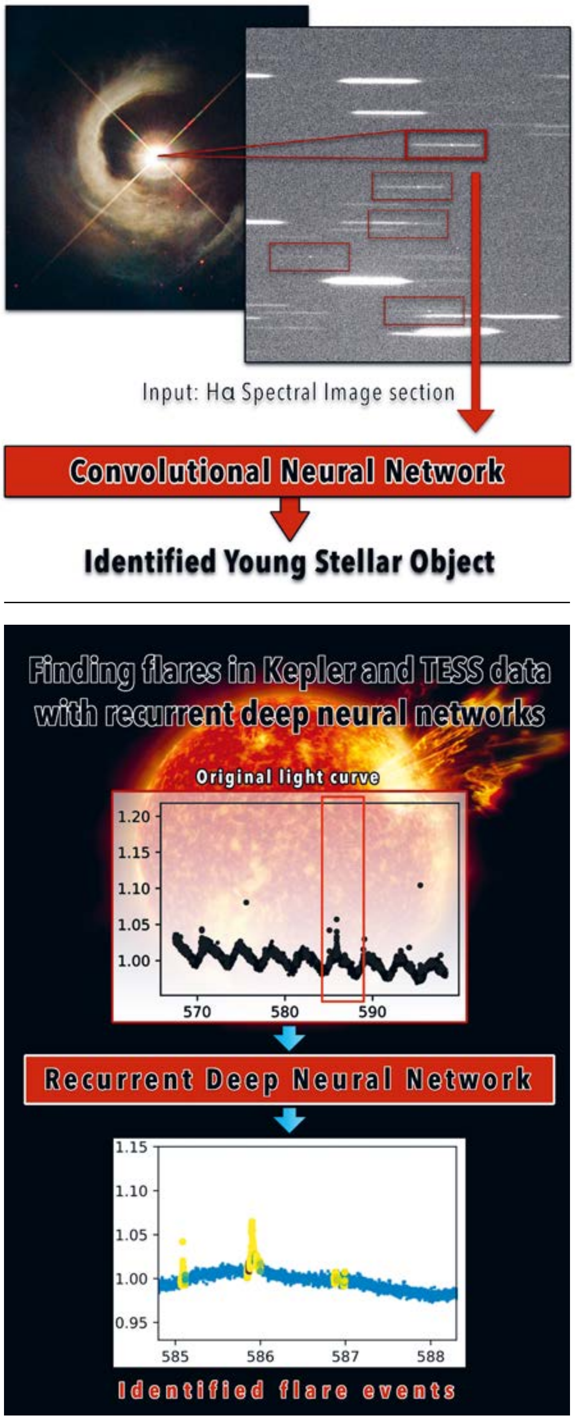


FIG. 1: Schematics of the image-based classification of periodic variable stars [5].

[2] https://www.lsst.org/

[3] https://www.skatelescope.org/

Our research team has developed high-precision algorithms to classify the light curves of periodic variable stars into a few main classes. If humans (professional astronomers) see a light curve of adequate quality, they can assign a variable type to the light curve. If the period is also known, then the classification can be close to perfect. However, human beings can classify only a few dozens of light curves (at maximum) every minute, and they cannot sustain this rate for long. Our supervised image-based machine learning method supplemented with numerical parameters (*e.g.*, period, luminosity, *etc.*) can reach or exceed the accuracy of human classification. Not to mention its speed that can be orders of magnitude faster than that of the human classifiers. Our neural network is a 'Multiple-Input Neural Network', which is made up of a Convolutional Neural Network and Multilayer Perceptrons and it can handle both image and numerical data. Recently we extended this method to incorporate sub-classes of variable stars, as well, with similar performance. Now we can not only tell cats and dogs apart but can classify millions of variable stars in a few minutes. And we don't need an army of graduate students to do the boring job. Although the job had had to be done manually at least once – our method needs labeled data, and labeled data are almost never available in astronomy for obvious reasons.



▸ **FIG. 2:** Upper panel: Identification of young stellar objects with a Convolutional Neural Network via detection of Hα emission lines (small point-like blobs) in low-resolution spectra. Lower panel: identification of stellar flares with a Recurrent Deep Neural Network.

Object detection – how to separate the wheat from the chaff?

Like many other disciplines, astronomy relies heavily on the possibilities offered by imaging tools. The large sky survey programs produce vast amounts of data – many terabytes per night – which cannot be processed manually on human time scales. With the use of machine learning tools, we are able to identify and categorise the distant galaxies of the Universe, stellar streams (remnants of dwarf galaxies) in the halo of our Galaxy or even the comets and asteroids of our own Solar System. Another application is the detection of young stellar objects (YSO) in astronomical recordings. These types of celestial bodies are stars in an early evolution stage, for example protostars or pre-main-sequence stars. Although these objects appear as "normal" stars in a CCD image, if we take low-resolution spectra, the YSOs' spectra will show strong emission in the Hα line. Based on this idea we created a Convolutional Neural Network. After proper training, the neural network can distinguish ordinary stars from YSOs with high precision.

How to turbo-boost your simulations with machine learning?

Large cosmological simulations are extremely resource intensive, since they take into account the action of gravity of billions of particles that trace the cosmic matter distribution including dark matter. This can easily be a bottleneck despite the ever-increasing computational capacity of the largest supercomputers. Well, who said that it is easy to simulate the whole universe? However, generative adversial networks (GANs) may come to the rescue [6]. This machine learning tool can generate cosmic web simulations that are quantitatively and qualitatively practically indistinguishable from real simulations, especially on large scales. The difference in computational time is huge: a fraction of a second for

[2] https://exoplanets.nasa.gov/

the GANs corresponds to many hours using traditional full-scale simulations.

Cosmology is not the only subfield where complex simulations are required. Much less demanding computations – at least in terms of mass particles involved –, but still very long integration times arise in celestial mechanics, for example when one has to decide whether a planetary system is stable or not on a timescale of billions of years. The discovery of close to 5000 exoplanets[4] during the last 28 years makes this problem even more acute. Direct integration of the motion of multiplanetary systems in large numbers is still prohibitive. However, machine learning can speed up the process by 5(!) orders of magnitude [7] by learning relevant, physically motivated features (chaos indicators, strengths of mean motion resonances, variance in eccentricity difference, *etc.*) from the beginning of the simulation period. That way the method can make accurate predictions about the stability properties of the system. ■

About the Authors

Róbert Szabó is the director of Konkoly Observatory of the Research Centre for Astronomy and Earth Sciences, which is the largest astronomical institute in Hungary. He specializes in observations and modelling of pulsating variables stars, Galactic archeology, exoplanets, and large sky surveys.

Attila Bódi is a postdoc researcher at Konkoly Observatory, working in R. Szabó's research group. He works on variable stars, machine learning, and classification methods.

Tamás Szklenár is a software developer working with R. Szabó's group. His main interest is machine learning methods and their applications. He is also interested in binary stars.

Acknowledgements

This project has been supported by the Lendület Program of the Hungarian Academy of Sciences, project No. LP2018-7/2021.

References

[1] M. V. Pruzhinskaya *et al.*, *MNRAS* **489**, 3591 (2019)

[2] Y. Ichinohe and S. Yamada, *MNRAS* **487**, 2874 (2019)

[3]Y. Sarkar *et al.*, *MNRAS* **510**, 6022 (2022)

[4] H. S. Chan *et al.*, ApJ submitted, arXiv:2112.03306

[5] T. Szklenár *et al.*, *ApJL* **897**, L12 (2020)

[6] A. C. Rodríguez *et al.*, *Comp. Astrophysics and Cosmology* **5**, 4 (2018)

[7] D. Tamayo *et al.*, *PNAS* **117**, 18194 (2020)