Exploring the Universe with Artificial Intelligence (UniversAI)
Proceedings IAU Symposium No. 397, 2026
I. Liodakis & V. Efthymiou, eds.
doi:10.1017/S174392132510478X

The "Variable" Universe with the *Gaia* mission and AI methods

L. Eyer[1], P. Huijse[2], N. Chornay[3], J. De Ridder[2], B. Holl[1,3], L. Rimoldini[3], K. Nienartowicz[3,4] and G. Jevardat de Fombelle[3]

[1]Department of Astronomy, University of Geneva, Chemin Pegasi 51, 1290 Versoix, Switzerland
email: laurent.eyer@unige.ch

[2]Institute of Astronomy, KU Leuven, Celestijnenlaan 200D, 3001 Leuven, Belgium

[3]Department of Astronomy, University of Geneva, Chemin d'Ecogia 16, 1290 Versoix, Switzerland

[4]Sednai sàrl, 4 Rue de Marbiers, 1204, Geneva, Switzerland

Abstract. The *Gaia* mission has observed over 2 billion stars repeatedly across the entire sky over 10 years, revealing the many astronomical objects that vary on human timescales from seconds to years. Its repeated astrometric, photometric, spectrophotometric and spectroscopic measurements create an unprecedented dataset to probe the variable celestial sources down to $G \approx 21$ mag. To extract meaningful results from these many time series for so many sources, we have used machine learning techniques for crossmatching, variability detection, and variability classification. This approach has now led to the largest catalogue of classified variable sources ever produced over the entire celestial sphere.

Keywords. stars: variables, methods: data analysis, catalogs, surveys

1. Introduction

Time-domain astronomy is an extremely rich and diverse field, yet we can still be amazed that so many celestial phenomena reveal themselves on human time scales. Among the many time-domain projects, the *Gaia* mission (Gaia Collaboration, Prusti et al. 2016) provides remarkable insights into the "variable sky". *Gaia* is a cornerstone mission of the Horizon 2000+ program of the European Space Agency. The spacecraft was launched in December 2013 and was passivated in March 2025. *Gaia* is unique in that, on a single platform and nearly simultaneously, it collected astrometric, photometric, spectrophotometric, and spectroscopic data across the entire celestial sphere, over a period of more than 10 years. Furthermore, thanks to its astrometric time series, the celestial sphere gets a geometrical depth!

So far, there have been three major data releases: DR1 in 2016 (Gaia Collaboration, Brown et al. 2016), DR2 in 2018 (Gaia Collaboration, Brown et al. 2018), and EDR3/DR3 in 2020/2022 (Gaia Collaboration, Brown et al. 2021; Gaia Collaboration, Vallenari et al. 2023). In addition, a focused release in 2023 included five articles, one of which was dedicated to the radial velocities of long-period variables (Gaia Collaboration, Trabucchi et al. 2023). Two further releases are planned: DR4 at the end of 2026 and DR5 in 2030.

Gaia has been an enormous success: its data are used across so many fields of astrophysics, and its publications break citation records (Parmar et al. 2024).

© The Author(s), 2026. Published by Cambridge University Press on behalf of The International Astronomical Union. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S174392132510478X Published online by Cambridge University Press

The vast number of sources (more than 2 billion) and the diversity of variable phenomena make automated machine-learning approaches indispensable. In this article, we focus specifically on the use of machine-learning methods for classifying variability in *Gaia* time-series data. The processing of *Gaia* data is iterative: with each new release, the data volume grows, calibrations improve, and the outputs become more diverse and complex. As a result, the development of methods and software has been a continuous effort.

The understaking to classify variable sources within *Gaia* has its roots in the *Hipparcos* mission (ESA 1997; Eyer 1998; Waelkens et al. 1998; Aerts, Eyer, & Kestens 1998), with a mixture of procedural approach rooted in astrophysical knowledge and also a multivariate discriminant analysis. Exploratory work on clustering techniques was carried out by Eyer & Blake (2005) on ASAS data, using k-means applied to fundamental variability features such as the period, amplitude, and the parameters of a Fourier series. Other methods were explored for *Gaia*; e.g. the stripe 82 of SDSS multiband photometry was analysed with a principle component analysis (Süveges et al. 2012). But a crucial step was achieved when the *Hipparcos* dataset was used to design classification methods for *Gaia*. Random Forest algorithms (Breiman 2001) proved particularly effective (Dubath et al. 2011). In this latter article, a comparison with a hierarchical classification approach developed by Blomme et al. (2011) for periodic variables was made and confirmed the effectiveness of the Random Forest approach. Rimoldini et al. (2012) applied Random Forest method to the "unsolved" variable stars of the *Hipparcos* mission and showed again its pertinence. Today, *Gaia* itself serves as a sandbox for developing and testing new methods. The latest efforts include the use of autoencoders (see Section 3). A particular feature of *Gaia* processing is that variability classification is first performed globally, after which specific classes are passed on for more detailed studies: for example, Cepheids and RR Lyrae stars. This creates a feedback loop between the machine-learning classifications and the evaluations of expert astronomers. Furthermore, for all classes, efforts were made to estimate completeness and contamination.

2. Supervised Classification

A central task of our activity in the *Gaia* Data Processing and Analysis Consortium (DPAC, Mignard et al. 2008) is given the *Gaia* time domain data to classify the source into variability types. Supervised classification was applied to *Gaia* Data from the first data release, then the second and the third. The method applied, as mentioned above, was based on feature extraction followed by the use of random forest classifiers. A total of 28 features were selected (from a much larger set) in the classification, some derived from periodograms and others from statistical measures; the full list is provided in Rimoldini et al. (2023). In addition, we compiled data from the literature to construct the training sets. As the adage goes, in supervised classification: garbage in, garbage out. The quality of the training set is therefore crucial. Thanks to the *Gaia* data, we can investigate variable sources reported in the literature. Typical problems with incorrectly labelled data include a variability level inconsistent with *Gaia* measurements, a light-curve shape incompatible with the assigned variability type, implausible *Gaia* colours and, when parallax is available, an incorrect position in the Hertzsprung-Russell diagram. The analysis of the published literature has often been of varying quality, with some notable positive exceptions, most prominently the OGLE project Udalski et al. (1992).

At every data release, the number of classified sources and the number of variability types are larger. DR1 contained 3,194 sources in two classes (Cepheids and RR Lyrae classes) for a limited region in the sky (part of the Large Magellanic Cloud). This release was more a showcase, a teaser for what is to come (Eyer et al. 2017). DR2 had 550,737

https://doi.org/10.1017/S174392132510478X Published online by Cambridge University Press

variable stars classified in 9 variability (sub)types (Holl et al. 2018; Rimoldini et al. 2019). DR3: listed 10.5 million variables in total into 24 † variability (sub)types, including 9.5 million variable stars and 1 million variable AGNs. Moreover, *Gaia* has made these classifications possible with nearly simultaneous measurements in astrometry, photometric time series, spectrophotometry, spectroscopy, and radial velocities (the last two for brighter sources) (Eyer et al. 2023; Rimoldini et al. 2023). The photometric time series are available, the spectrophotometry was used to classify carbon stars and an unprecedented sample of radial velocities time series was published (Lebzelter et al. 2023). The photometric and radial velocity time series are publicly available in the *Gaia* ESA archive‡.

One characteristic of working within an ESA consortium is that we are bound to a tight schedule. Processing time is often a bottleneck. Although, as Donald Knuth famously remarked, "premature optimisation is the root of all evil", we sometimes need to make non-optimal choices to reduce the time required to meet deadlines. In recent years, a major game changer has been the use of GPUs (Fluke 2012), which have significantly accelerated many time-consuming computations, such as period-search methods.

3. Variational Autoencoders

Autoencoders are neural network methods that reduce the dimensionality of input data (Goodfellow, Bengio & Courville 2016). An encoder compresses the input (for example, an image or a time series) into a lower-dimensional representation, and a decoder reconstructs the input from this compressed form. By training the network to minimise the difference between the reconstruction and the original, the autoencoder learns a reduced but meaningful description of the data. The compressed quantities are referred to as the latent variables.

For DR4, we will provide a separate catalogue output based on variational autoencoders (VAEs), building on work carried out with *Gaia* DR3 data (Huijse et al. 2025). In that study, three VAEs were trained, each with five latent variables, on different *Gaia* data products (1) the BP and RP mean spectrophotometry, (2) the folded G light curve, and (3) the DMDT (magnitude differences versus the corresponding time pairs) representation of the light curve (Mahabal et al. 2017). The folded light curve is especially informative for periodic sources, while the DMDT method better captures aperiodic variability. From these three VAEs, 15 latent variables were obtained. After normalising the inputs, three additional features were added: the log of the dominant frequency, the standard deviation and the colour. Together, these 18 features provide a compact yet physically motivated description of each source. Importantly, parallax (and thus distance) was not included.

A second-stage VAE was then applied to reduce these features to two latent variables, yielding a 2D projection of the data. This projection successfully groups the different major classes and shows strong correlation with astrophysical properties, demonstrating the method's effectiveness. Huijse et al. (2025) also showed that the latent representation can serve as input to other machine learning methods, enabling:

- Anomaly detection, when combined with techniques such as local outlier factor or isolation forests.
- Clustering, using methods such as k-means or DBSCAN (Density-Based Spatial Clustering of Applications with Noise, Ester et al. 1996).
- Classification, in combination with algorithms such as k-NN or random forest.

† In fact, there are 24+1 variable types, because the supervised classification was able to identify 2.5 million galaxies due to spurious variability caused by the extended nature of these objects and the way Gaia acquired these data.

‡ https://gea.esac.esa.int/archive/

https://doi.org/10.1017/S174392132510478X Published online by Cambridge University Press

The forthcoming DR4 analysis will allow us to fully assess the potential of this approach.

4. Classification with GaiaVari, a citizen science project

GaiaVari† is a citizen science project on the Zooniverse platform aimed at classifying subsets of variable sources from *Gaia* DR3 (Eyer et al. 2023; Eyer 2024) The classification relies on fundamental astronomical plots, such as time series, folded light curves, HR/colour-magnitude diagrams, and location in the Milky Way. We believe it is essential to promote high-level interaction not only between machine-learning experts and astronomers but also with the wider public. Such an activity demonstrates the fundamental importance of visualising properties of individual sources. This does not mean that having many sources makes it pointless to look at just a few. On the contrary, the more sources you have, the more meaningful subsamples you can create!

5. Conclusion and perspective

For the variable sources, we have been preparing the data processing and analysis of the *Gaia* mission for many years within the *Gaia* DPAC Consortium, testing and applying different machine-learning methods. Our goal has been to produce catalogues with a holistic approach that can be used by the scientific community and amateur astronomers. The result is a significant leap forward: at the time of the third data release publication, we provided, across the entire sky, the largest and most homogeneous set of variable sources to date, breaking many records in the number of many variability types. Our analysis provides, for the first time, a comprehensive view of variability among Milky Way stars across the Colour-Absolute Magnitude Diagram Gaia Collaboration et al. (2019).

This is not the end of the road for the *Gaia* data and associated methods to analyse them. Although the spacecraft terminated its data collection, future releases will include iteratively larger datasets, and machine learning techniques are advancing rapidly. In DR4, the time series used for analysis will be twice as long as in DR3, and it will double again in DR5. Also note that from DR4 all time series will be available to the public. From other surveys, we will be able to benefit from OGLE, ZTF and TESS data for the training sets and validation.

Within the *Gaia* consortium, each Data Release is an opportunity to test new methods. Indeed, *Gaia* provides a valuable testbed for novel machine-learning approaches, with the published catalogues serving as benchmarks against which new techniques can be compared.

In time-domain astronomy, the next major leap forward will likely come from the Vera Rubin Observatory (LSST; LSST Science Collaboration et al. 2009), which will explore the faint variable sky from the ground using a multi-band photometric system. Combining the strengths of both *Gaia* and Rubin will enable significant progress in the field over the coming years.

References

Aerts C., Eyer L., Kestens E., 1998, A&A, 337, 790

Blomme J., Sarro L. M., O'Donovan F. T., Debosscher J., Brown T., Lopez M., Dubath P., et al., 2011, MNRAS, 418, 96. doi:10.1111/j.1365-2966.2011.19466.x

Breiman L., 2001, MachL, 45, 5. doi:10.1023/A:1010933404324

Dubath P., Rimoldini L., Süveges M., Blomme J., López M., Sarro L. M., De Ridder J., et al., 2011, MNRAS, 414, 2602. doi:10.1111/j.1365-2966.2011.18575.x

ESA, 1997, ESASP, 1200

† https://www.gaiavari.space

https://doi.org/10.1017/S174392132510478X Published online by Cambridge University Press

Ester M., Kriegel H.-P., Sander J., Xu X., 1996, in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD'96 (AAAI Press), 226–231
Eyer L., 1998, PhDT
Eyer L., Blake C., 2005, MNRAS, 358, 30. doi:10.1111/j.1365-2966.2005.08651.x
Eyer L., Mowlavi N., Evans D. W., Nienartowicz K., Ordonez D., Holl B., Lecoeur-Taibi I., et al., 2017, arXiv, arXiv:1702.03295. doi:10.48550/arXiv.1702.03295
Eyer L., Audard M., Holl B., Rimoldini L., Carnerero M. I., Clementini G., De Ridder J., et al., 2023, A&A, 674, A13. doi:10.1051/0004-6361/202244242
Eyer L., Nienartowicz K., Garcia Lario P., Ratajczak M., Pomierny J., Pingeon G., Audard M., 2023, eas..conf, 1738
Eyer L., 2024, ees..conf, 27. doi:10.48550/arXiv.2412.02744
Fluke C. J., 2012, ASPC, 461, 3. doi:10.48550/arXiv.1111.5081
Gaia Collaboration, Brown A. G. A., Vallenari A., Prusti T., de Bruijne J. H. J., Mignard F., Drimmel R., et al., 2016, A&A, 595, A2. doi:10.1051/0004-6361/201629512
Gaia Collaboration, Brown A. G. A., Vallenari A., Prusti T., de Bruijne J. H. J., Babusiaux C., Bailer-Jones C. A. L., et al., 2018, A&A, 616, A1. doi:10.1051/0004-6361/201833051
Gaia Collaboration, Brown A. G. A., Vallenari A., Prusti T., de Bruijne J. H. J., Babusiaux C., Biermann M., et al., 2021, A&A, 649, A1. doi:10.1051/0004-6361/202039657
Gaia Collaboration, Eyer L., Rimoldini L., Audard M., Anderson R. I., Nienartowicz K., Glass F., et al., 2019, A&A, 623, A110. doi:10.1051/0004-6361/201833304
Gaia Collaboration, Prusti T., de Bruijne J. H. J., Brown A. G. A., Vallenari A., Babusiaux C., Bailer-Jones C. A. L., et al., 2016, A&A, 595, A1. doi:10.1051/0004-6361/201629272
Gaia Collaboration, Trabucchi M., Mowlavi N., Lebzelter T., Lecoeur-Taibi I., Audard M., Eyer L., et al., 2023, A&A, 680, A36. doi:10.1051/0004-6361/202347287
Gaia Collaboration, Vallenari A., Brown A. G. A., Prusti T., de Bruijne J. H. J., Arenou F., Babusiaux C., et al., 2023, A&A, 674, A1. doi:10.1051/0004-6361/202243940
Goodfellow I., Bengio Y., Courville A., 2016, Deep Learning. MIT Press, Cambridge, MA. ISBN: 0262035618
Holl B., Audard M., Nienartowicz K., Jevardat de Fombelle G., Marchal O., Mowlavi N., Clementini G., et al., 2018, A&A, 618, A30. doi:10.1051/0004-6361/201832892
Huijse P., De Ridder J., Eyer L., Rimoldini L., Holl B., Chornay N., Roquette J., et al., 2025, arXiv, arXiv:2505.16320. doi:10.48550/arXiv.2505.16320
Lebzelter T., Mowlavi N., Lecoeur-Taibi I., Trabucchi M., Audard M., García-Lario P., Gavras P., et al., 2023, A&A, 674, A15. doi:10.1051/0004-6361/202244241
LSST Science Collaboration, Abell P. A., Allison J., Anderson S. F., Andrew J. R., Angel J. R. P., Armus L., et al., 2009, arXiv, arXiv:0912.0201. doi:10.48550/arXiv.0912.0201
Mahabal A., Sheth K., Gieseke F., Pai A., Djorgovski S. G., Drake A., Graham M., et al., 2017, arXiv, arXiv:1709.06257. doi:10.48550/arXiv.1709.06257
Mignard F., Bailer-Jones C., Bastian U., Drimmel R., Eyer L., Katz D., van Leeuwen F., et al., 2008, IAUS, 248, 224. doi:10.1017/S1743921308019145
Parmar A., Bonnet R.-M., De Marchi G., García-Lario P., Kuulkers E., Pilbratt G., Sánchez-Fernández C., et al., 2024, ISSI Scientific Report Series 18, Springer, Cham, doi:10.1007/978-3-031-69004-4
Rimoldini L., Dubath P., Süveges M., López M., Sarro L. M., Blomme J., De Ridder J., et al., 2012, MNRAS, 427, 2917. doi:10.1111/j.1365-2966.2012.21752.x
Rimoldini L., Holl B., Audard M., Mowlavi N., Nienartowicz K., Evans D. W., Guy L. P., et al., 2019, A&A, 625, A97. doi:10.1051/0004-6361/201834616
Rimoldini L., Holl B., Gavras P., Audard M., De Ridder J., Mowlavi N., Nienartowicz K., et al., 2023, A&A, 674, A14. doi:10.1051/0004-6361/202245591
Udalski A., Szymanski M., Kaluzny J., Kubiak M., Mateo M., 1992, AcA, 42, 253
Süveges M., Sesar B., Váradi M., Mowlavi N., Becker A. C., Ivezić Ž., Beck M., et al., 2012, MNRAS, 424, 2528. doi:10.1111/j.1365-2966.2012.21229.x
Waelkens C., Aerts C., Kestens E., Grenon M., Eyer L., 1998, A&A, 330, 215

https://doi.org/10.1017/S174392132510478X Published online by Cambridge University Press