A&A, 693, A268 (2025)
https://doi.org/10.1051/0004-6361/202452429

Astronomy & Astrophysics

Variability in hot sub-luminous stars and binaries: Machine-learning analysis of *Gaia* DR3 multi-epoch photometry

P. Ranaivomanana[1,2,⋆], M. Uzundag[2], C. Johnston[1,2,3], P. J. Groot[1,4,5,6], T. Kupfer[7,8], and C. Aerts[1,2,9]

[1] Department of Astrophysics/IMAPP, Radboud University, PO Box 9010, 6500 GL Nijmegen, The Netherlands
[2] Instituut voor Sterrenkunde, KU Leuven, Celestijnenlaan 200D, 3001 Leuven, Belgium
[3] Max-Planck-Institut für Astrophysik, Karl-Schwarzschild-Straße 1, 85741 Garching bei München, Germany
[4] Department of Astronomy, University of Cape Town, Private Bag X3, Rondebosch 7701, South Africa
[5] South African Astronomical Observatory, PO Box 9, Observatory 7935, South Africa
[6] The Inter-University Institute for Data Intensive Astronomy, University of Cape Town, Private Bag X3, Rondebosch 7701, South Africa
[7] Hamburger Sternwarte, University of Hamburg, Gojenbergsweg 112, 21029 Hamburg, Germany
[8] Texas Tech University, Department of Physics & Astronomy, Box 41051, 79409 Lubbock, TX, USA
[9] Max Planck Institute for Astronomy, Königstuhl 17, 69117 Heidelberg, Germany

Received 30 September 2024 / Accepted 3 December 2024

ABSTRACT

Context. Hot sub-luminous stars represent a population of stripped and evolved red giants that is located on the extreme horizontal branch. Since they exhibit a wide range of variability due to pulsations or binary interactions, it is crucial to unveil their intrinsic and extrinsic variability to understand the physical processes of their formation. In the Hertzsprung-Russell diagram, they overlap with interacting binaries such as cataclysmic variables (CVs).
Aims. By leveraging the most recent clustering algorithm tools, we investigate the variability of 1576 candidate hot subdwarf variables using comprehensive data from *Gaia* DR3 multi-epoch photometry and Transiting Exoplanet Survey Satellite (TESS) observations.
Methods. We present a novel approach that uses the t-distributed stochastic neighbour embedding and the uniform manifold approximation and projection dimensionality reduction algorithms to facilitate the identification and classification of different populations of variable hot subdwarfs and CVs in a large dataset. In addition to the publicly available *Gaia* time-series statistics table, we adopted additional statistical features that enhanced the performance of the algorithms.
Results. The clustering results led to the identification of 85 new hot subdwarf variables based on *Gaia* and TESS light curves and of 108 new variables based on *Gaia* light curves alone, including reflection-effect systems, HW Vir, ellipsoidal variables, and high-amplitude pulsating variables. A significant number of known CVs (140) distinctively cluster in the 2D feature space among an additional 152 objects that we consider candidates for new CVs.
Conclusions. This study paves the way for more efficient and comprehensive analyses of stellar variability from ground- and space-based observations, and for the application of machine-learning classifications of candidate variable stars in large surveys.

Key words. methods: data analysis – methods: statistical – techniques: photometric – surveys – subdwarfs – stars: variables: general

1. Introduction

Hot sub-luminous stars are hot and compact evolved low-mass stars that are located on the extreme horizontal branch, between the main sequence (MS) and the white dwarf sequence (Heber 2009, 2016). In a Hertzsprung-Russell diagram (HRD), they occupy B and O spectral types and form the population of hot subdwarf B (sdB) and O (sdO) stars. A recent study of a 500 pc volume-limited sample of hot sub-luminous stars reported that they are dominated by the sdB population (∼60%; Dawson et al. 2024). Most of this population are thought to have a canonical core mass of 0.47 $M_\odot$ and thin hydrogen layers ($\sim 10^{-4}$–10^{-2} $M_\odot$; Saffer et al. 1994; Brassard et al. 2001). Their thin envelope mass suggests that sdBs are the remnant cores of low-mass red giant stars that were stripped through binary interactions, which introduced a different evolutionary path than for normal horizontal branch stars. This envelope mass prevents them from supporting H-shell burning. After depletion of helium in the sdB cores, on a timescale of $\sim 10^8$ yr (Dorman et al. 1993; Ostrowski et al. 2021), they first become sdOs and then evolve to the white dwarf cooling stage.

Evolutionary calculations showed that sdB progenitors likely underwent binary interactions (Han et al. 2002, 2003), including common-envelope ejection (CEE; for short-period binaries with a period of 0.1–10 days), stable Roche-lobe overflow (RLOF; for long-period or composite binaries with periods of 450–1600 days; Vos et al. 2020), and mergers (e.g. He white dwarf + He white dwarf; Webbink 1984). Observational studies corroborated this (Pelisoli et al. 2020), and multiple studies reported a significant fraction of hot subdwarfs in binary systems, either in close binaries with a MS or white dwarf companion (e.g. Geier et al. 2022; Schaffenroth et al. 2022, 2023), or in wide binaries with cool MS companions (e.g. Deca et al. 2012; Vos et al. 2019, 2020). This diversity makes them an excellent population for studying binary star evolution. In addition, a broad range of unseen companions have been confirmed to exist in hot subdwarfs, such as low-mass MS stars (dM), brown dwarfs, and white dwarf companions (Kupfer et al. 2015; Geier et al. 2010, 2022, 2023) through the project called Massive

⋆ Corresponding author; princy.ranaivomanana@ru.nl

Unseen Companions to Hot Faint Underluminous Stars from the Sloan Digital Sky Survey (MUCHFUSS). The existence of these companions and their nature is often shown by the behaviour of photometric light curves of the hot subdwarfs, such as ellipsoidal variability for white dwarf companions and reflection effects for low-mass companions (Schaffenroth et al. 2022; Barlow et al. 2022).

A population of hot subdwarfs was also found to exhibit pulsations, and asteroseismology was used to study their structure and evolution (e.g. Charpinet et al. 2010; Van Grootel et al. 2010; Reed et al. 2020; Sahoo et al. 2020; Silvotti et al. 2022; Krzesinski & Balona 2022; Uzundag et al. 2021, 2023, 2024). While the mechanism for exciting pulsations in subdwarfs is thought to be understood (i.e. the κ-mechanism operating on the Fe opacity bump; Charpinet et al. 1997; Fontaine et al. 2003), it is unclear why only a handful of subdwarfs are observed to pulsate while most do not. Theoretical work has demonstrated that atomic diffusion is required, but it is unclear whether other aspects such as the evolution history of the binary also play a role (Hu et al. 2008, 2011; Bloemen et al. 2014).

It is essential to increase the detection of new variable hot subdwarfs to enable a robust characterisation of their variability and to improve our understanding of these stars. In addition to spectroscopic identifications of hot subdwarfs (e.g. Luo et al. 2019; Lei et al. 2020, 2023), which are often observationally expensive, previous efforts to identify candidate hot subdwarfs were made mainly based on their locations in the colour-magnitude diagram and proper motion selection criteria (Geier et al. 2019; Geier 2020) using *Gaia* DR2 observations (Gaia Collaboration 2018). Following similar steps, Culpan et al. (2022) compiled a large catalogue of more than 60 000 confirmed and candidate hot subdwarfs observed from *Gaia* EDR3 data (Gaia Collaboration 2021a). These selections are frequently affected by contamination from low-mass MS stars, cataclysmic variables (CVs), and white dwarfs (Geier et al. 2019; Culpan et al. 2022; Barlow et al. 2022). Given this contamination, it is critical for target selections to develop an effective framework to separate hot subdwarfs from other populations of blue objects in the HR diagram and characterise their variability in multiple time-domain surveys.

The interest in developing machine-learning algorithms to automate the variability search and characterisation of time-series data in time-domain astronomy has been strong (e.g. Kim et al. 2021; Cui et al. 2022; Eyer et al. 2023; Monsalves et al. 2024) due to the growing volume of data generated by large surveys, such as the All Sky Automated Survey (ASAS; Pojmanski 2002), the *Zwicky* Transient Facility (ZTF; Bellm et al. 2019), and the *Gaia* mission (Gaia Collaboration 2023). As the majority of these algorithms either depend on a particular survey (e.g. based in space or on the ground) or are task-oriented (e.g. a planet transit detection), their application is often limited to a certain number of specific cases and goals.

To remedy this, we present a machine-learning framework for identifying variable hot subdwarfs and CVs based on photometric time series alone. Our methods can be broadly applied to any photometric data, such as those from the BlackGEM (Groot et al. 2024), the Gravitational-wave Optical Transient Observer (GOTO; Steeghs et al. 2022), and the Legacy Survey of Space and Time (VRO/LSST; Ivezić et al. 2019) missions. The structure of this paper is as follows: In Sect. 2 we describe the data and methods. This is followed by feature engineering and the cluster analysis in Sect. 3. The results of the variability classification are provided and discussed in Sect. 4. Our conclusion and future prospects are presented in Sect. 5.

2. Data and methods

2.1. *Gaia* observations

The precise astrometric and photometric measurements provided by *Gaia* significantly boost the identification of the population of candidate hot subdwarfs in the colour-magnitude diagram. Culpan et al. (2022) compiled a catalogue of 61 585 candidate hot subdwarfs based on colour, absolute magnitude, and reduced proper motion selection criteria in *Gaia* EDR3 (Gaia Collaboration 2021a), which served as the basis of this work. The release of *Gaia* DR3 multi-epoch photometry (Eyer et al. 2023) allowed us to cross-match this catalogue to find candidates with available light curves and further study their variability. This resulted in 2114 objects with available epoch photometry using the *Gaia* flag has_epoch_photometry = True. The remaining 59 471 objects were excluded from the analysis because no *Gaia* light curves are available for them.

Using the *Gaia* datalink service and the *astroquery.Gaia* package (Ginsburg et al. 2019), we extracted the light curves of these objects in the three *Gaia* filter bands (*G*, *BP*, and *RP*). Before we searched for periodicity, we preliminarily assessed the quality. First, we retained objects with reliable parallax measurements (parallax_over_error > 5). Second, the *Gaia* boolean quality flag `reject_by_variability` was used to remove data points rejected by the *Gaia* variability pipeline (Eyer et al. 2023), and then objects with at least 25 observations in any of the three band light curves were selected, following the minimum number of observations suggested by Morales-Rueda et al. (2006) for detecting stellar variability. For the *Gaia* astrometric quality control, known as the re-normalised unit weight error (RUWE), RUWE < 7 was adapted as a substantial number of spectroscopically identified hot subdwarfs were observed to exceed the recommended RUWE < 1.4 limit up to RUWE = 7 (see Dawson et al. 2024 for more details). These selections resulted in 1682 light curves that were ready for analysis. Their *Gaia* *G*-band light curves have a typical median signal-to-noise (S/N) ratio estimate (standard deviation of the magnitudes over the rms of the magnitude uncertainties) of 3.5 and a median number of observations of about 40, as well as a median magnitude of ∼15 mag.

2.2. Frequency analysis

The population of hot subdwarfs hosts diverse types of variability, including pulsating variables and eclipsing binaries, from close- to wide-binary systems. Therefore, their variability exhibits a wide range of timescales from minutes to months and of morphologies from sharp eclipses to sinusoidal pulsations. Following the success of our frequency-search algorithms in finding dominant frequencies in multi-band, heteroscedastic, and irregularly sampled light curves of candidate hot subdwarfs from the MeerLICHT telescope (Ranaivomanana et al. 2023), we applied the same approach to search for periodicity in *Gaia* light curves. In brief, this method combines Fourier-based calculations, namely the generalised Lomb-Scargle periodogram, and phase-dispersion measurements, known as Lafler-Kinman statistics, to alleviate the effects of noise and data gaps in a periodogram. This hybrid approach is referred to as the Ψ-static (Saha & Vivas 2017), where the notation Ψ is used to represent the periodogram throughout this work.

In the frequency grid search, the search was performed from zero up to 360 day^{-1} according to the Nyquist-frequency of the 2-minute cadence of the Transiting Exoplanet Survey Satellite (TESS; Ricker et al. 2015) short-cadence observations, which

were used for comparison with the *Gaia* variability in Sect. 4. The frequency step was finely tuned and was defined as the inverse of the total time base divided by an oversampling factor of 10, following results in the literature that showed that this value is appropriate to ensure that no dominant frequency peaks are missed and to prevent a poor period estimation, which would occur if its value were taken too low (VanderPlas 2018; Schwarzenberg-Czerny 1996). In addition, the dominant frequency we found was further optimised by fine-tuning the frequency step with an oversampling factor of 100. This was only done in a small frequency window around the dominant frequency, where a frequency window size ten times larger than the original frequency step was used on either side of the peak.

2.3. Uncertainties in the frequency estimates

The uncertainties in the dominant frequencies were estimated by adopting a Monte Carlo approach, where the frequency algorithm was run 1000 times. The standard deviation of the dominant frequencies was taken as an estimate of the frequency uncertainty. Each iteration consisted of (1) drawing a sample from a normal distribution with zero mean and a width of the magnitude errors per observation, and (2) creating a new light curve by adding the sample to the original light curve. The magnitude errors in the original light curves were kept in the new light curves. The iterations finally consisted of (3) running the algorithm on the new light curve using the same fine-tuned frequency window as in the frequency optimisation. Due to the finite sampling step in the frequency grid, each iteration could result in the same identified dominant frequency. To mitigate this, the frequency grids were shifted by 1/1000th of the frequency step for each of the 1000 iterations to ensure that the frequency search was not confined to the same frequency peak in each iteration.

3. Feature engineering

3.1. Variability analysis

After we computed the dominant frequencies for all candidates, a robust, unbiased method was required to determine the significance of the peaks and measure the reliability of the variability. Although the false-alarm probability (FAP; Scargle 1982; Baluev 2008) was frequently used in the literature to measure the significance of the frequency peak, it is poorly adapted to variables in the high-frequency domain and in the case of signals with red noise (VanderPlas 2018). Additionally, the interpretation of the FAP becomes complex in our case, where two independent periodograms were combined in the hybrid approach. Therefore, we addressed this by exploring machine-learning clustering algorithms to distinguish candidates with different significance levels and variability.

We explored various summary statistics that are capable of unveiling the fidelity of the frequency peaks and the variability in the *Gaia* time series. It was necessary to extract these parameters from the data to work with the clustering algorithms described in the next sections. First, we extracted the *Gaia* variability summary statistics table[1], which consists of statistical parameters (54 in total, excluding boolean parameters and object IDs) that were computed using the *Gaia* DR3 time series (Eyer et al. 2023). Second, after normalising the maximum amplitude in the Ψ-periodogram to one, we computed additional statistical features (24 features) that were specifically designed to help us define the significance of the peak, such as the 95th percentile of the amplitude for the 100 peaks with the highest amplitude, the 99th percentile of the amplitudes for the full spectrum, and the number of frequencies with amplitudes above 0.5. These features were found to be useful for distinguishing objects with a clear variability, as we discuss in Sect. 3.2. We obtained 84 features in total (see Table A.1) when we combined these features with the *Gaia* statistics table and another six parameters from the *Gaia* DR3 source database (Gaia Collaboration 2023), such as *BP*−*RP*, parallax, and RUWE. Entries with missing values were removed from the table, which left 1576 final candidates out of the 1682 objects.

3.2. Dimensionality reduction

The next step was to transform these features into a lower-dimensional space such that we were able to visualise and identify possible clusters. This was done by applying dimensionality reduction techniques to our data, which convert high-dimensional features into a 2D feature space. It is common practice to reduce the dimensionality in machine learning, and it has been used extensively in astronomy to visualise and interpret data (Kao et al. 2024; Liao et al. 2024; Pantoja et al. 2022). We explored two non-linear dimensionality reduction techniques: the t-distributed stochastic neighbour embedding (t-SNE; van der Maaten & Hinton 2008) and the uniform manifold approximation and projection (UMAP; McInnes et al. 2018a). These were chosen over other techniques such as the principal component analysis (PCA) because they are able to find non-linear structures in data and are straightforward to implement.

Since our data had more than 80 features, it was important to remove highly correlated features that might lead to noise in the visualisation (Kuhn & Johnson 2019). This also helped the algorithms, notably t-SNE, to efficiently map the high-dimensional to low-dimensional space. To identify correlated features, we calculated the Pearson correlation coefficient between all pair-wise combinations of features and excluded one feature from each pair for correlation values above 0.95. This reduced the number of features to 49, which is also recommended[2] (∼50) to efficiently optimise the t-SNE algorithm.

We further ranked the features using a random forest algorithm (Breiman 2001), which is a commonly used technique for obtaining relative feature importance scores (e.g. Richards et al. 2012). The importance score of each feature is determined based on its ability to split the data into pure nodes (nodes with instances belonging to the same class) in the individual decision trees of the random forest model (see Breiman 2001 for more details). At this stage, the sole purpose was to obtain the feature scores. Therefore, the default random forest model hyperparameters (e.g. the number of estimators) were used to fit the data. Upon model fitting, we obtained the relative importance scores of each feature. These scores were used to optimise the t-SNE and UMAP algorithms in Sects. 3.2.1 and 3.2.2.

We also manually labelled each object based on their phase-folded diagrams, where objects that exhibited an obvious variability were labelled as 0, and those with an ambiguous variability were labelled as 1. These labels were used when fitting the random forest algorithm. In addition, labelling the data allowed us to examine the clustering performances and to visualise the physical or statistical distribution of each class (e.g.

[1] https://doi.org/10.17876/Gaia/dr.3/92

[2] https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

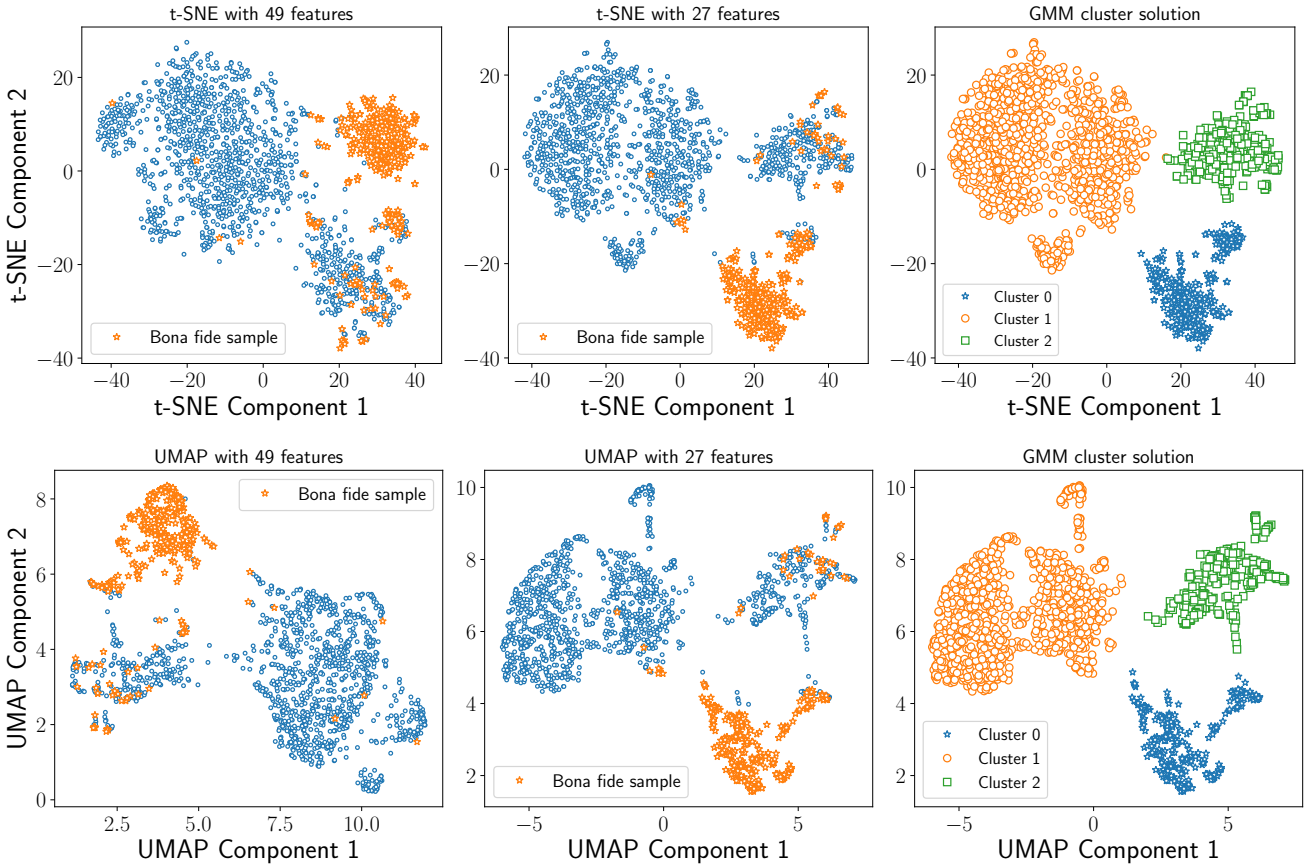


Fig. 1. Clustering results using the t-SNE (top panels) and UMAP (bottom panels) dimensionality reduction algorithms and clustering labels (right panels) from the Gaussian mixture model (GMM). The left and middle panels represent the 2D components using 49 and 27 features, respectively. The open orange stars in these panels correspond to the manually selected objects with clear variabilities.

period distribution) in the clusters shown in Fig. 1, which we discuss further throughout the paper.

3.2.1. Dimensionality reduction with t-SNE

We implemented the `TSNE` module from the `scikit-learn` Python library (Pedregosa et al. 2011), where two crucial parameters, namely perplexity and learning rate, were optimised, while the other parameters were kept to their default values. The perplexity can be seen as a tuning parameter that measures the effective number of nearest neighbours to be considered to construct the low-dimensional embedding. Before running the t-SNE algorithm, we first scaled each feature to have a zero mean and unit standard deviation, which helped the algorithm to be more efficient in finding structures in the data. The optimised values of the two parameters are perplexity = 50 and learning rate = 600. With these settings and the 49 features, Fig. 1 shows the transformed low-dimensional projections, where we can identify three main clusters, namely cluster 0, cluster 1, and cluster 2. These are discussed in more detail in Sect. 3.3. The open orange stars in the left and middle panels of Fig. 1 represent the objects we labelled manually, most of which belong to one cluster. To label these clusters, we fit the 2D projection data to a Gaussian mixture model (implemented in scikit-learn) with three mixture components. The advantage of using this model is that it provides the probability of each object to belong to a cluster. The quality of the class labels predicted by the Gaussian mixture model was evaluated using the so-called silhouette score (Rousseeuw 1987), in addition to a visual inspection of the graphical output. This evaluation metric compares how well data points match their designated cluster to other clusters. We obtained a silhouette score of 0.535, which is generally considered to indicate a reasonable clustering solution (i.e. >0.5; Rousseeuw 1987). We further improved this by iteratively removing the least important features from the importance scores computed above that might cause noise in the low-dimensional representation. In other words, we stopped the iterative process when no further improvements were visually detected in the output clusters and in the silhouette score. This resulted in 27 features with a silhouette score of 0.567. The t-SNE 2D representation of this result is shown in Fig. 1 together with the Gaussian mixture clustering solution. These 27 features are described in Table A.2 and are used throughout the rest of the analysis.

The dominant features for the manually labelled objects include the 95th percentile of the first 100 frequency peaks, the number of peaks above 0.5 of the normalised Ψ periodogram, and the 99th percentile of all periodogram peaks. However, they do not imply that these top features alone can explain the separation of the three clusters in the 2D feature space; it only means that their importance scores are higher than those for the rest of the features, as shown in Fig. A.2. As previously mentioned, the aim of dimensionality reduction algorithms is to build new low-dimensional features from linear or non-linear combinations of high-dimensional features while preserving as much of the original information as possible. Since the low-dimensional features are mixtures of the original ones, we cannot conclude from the

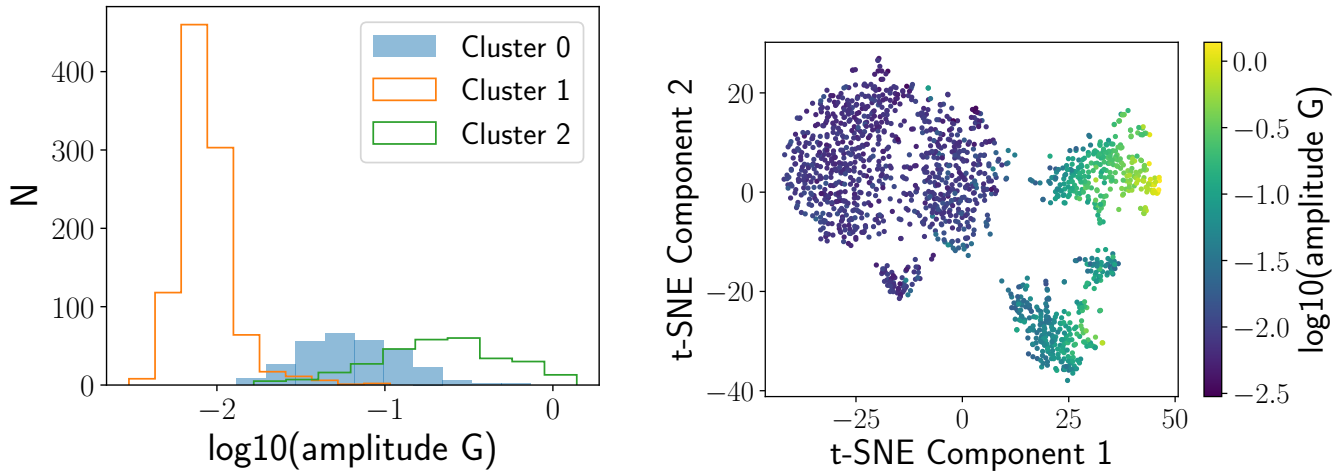


Fig. 2. Amplitude distribution of each cluster in the *Gaia G* band (left panel) and gradient of the variability amplitude in the *G* band across the t-SNE components (right panel).

2D representation that a specific or a group of a few features cause the distinction of the clusters.

3.2.2. Dimensionality reduction with UMAP

Similar to t-SNE, UMAP (McInnes et al. 2018a) is a non-linear algorithm for high-dimensional data visualisation, except that its approach to dimensionality reduction is grounded in manifold theory and topological data analysis rather than probabilistic modelling as in t-SNE. The UMAP algorithm is implemented in the umap-learn Python package (McInnes et al. 2018b). We ran the UMAP algorithm with its default parameter values and the selected 27 features in Sect. 3.2.1, which already resulted in reasonable silhouette score values and a distinctive visualisation (Fig. 1). The same features as obtained from t-SNE were also used when running the UMAP algorithm to show that both algorithms output the same results using the same features, and to obtain meaningful clustering results. The obtained silhouette scores are very similar for the 27 features (0.597) and 49 features (0.599). The cluster labels were again obtained from the Gaussian mixture model. We identified three main clusters similar to those found with the t-SNE algorithm, which confirms the existence of these clusters in our data. The next section compares the results from the two algorithms.

3.3. Cluster analysis and candidate selection

It is worth examining whether the three clusters found by t-SNE and UMAP represent the same objects. Of the t-SNE and UMAP components, 290 and 297 objects are part of cluster 0; 990 and 991 objects in cluster 1; and 296 and 288 objects in cluster 2, respectively. Therefore, we cross-matched the objects in the three clusters from both algorithms and found a total number of 1563/1576 matches (~99%): 289, 988, and 286 matches from cluster 0, cluster 1, and cluster 2, respectively. This shows that the two clustering approaches are highly consistent. We examined the 13 mismatched objects because the clustering results for t-SNE and UMAP matched for 1563 out of 1576 objects. Eight of these 13 objects belong to cluster 0 in UMAP and to cluster 2 in t-SNE. These objects exhibit large peak-to-peak magnitudes in the *Gaia G* band, with variations of at least 0.5 mag. In the 2D t-SNE plot, they are located near the border of cluster 2, close to cluster 0, which may explain the mismatch in the cluster labels between UMAP and t-SNE for these objects. The remaining 5 of the 13 objects either appear in cluster 1 in UMAP and cluster 2 in t-SNE, or vice versa, and they are similarly positioned at the borders of each cluster. We did not observe any peculiar objects in addition to these cases.

As our primary goal was to identify objects with significant and clear variability among the clusters, we visually examined the light curves of the objects in each cluster. We observed that the three clusters reflect the clarity of the light-curve variability, which can be translated into the light-curve S/N ratio. More precisely, cluster 1 contains objects with a dubious variability that might be related to light curves with a relatively low S/N ratio; cluster 2 primarily consists of objects with ambiguous light-curve shapes but high variability amplitudes; and cluster 0 is dominated by objects with a clear variability that is associated with high S/N ratio light curves. Some examples of light curves in each cluster are shown in Fig. A.1, where the top panels represent clear variables that are typical for cluster 0, the middle panels correspond to unclear variables found in cluster 1, and the bottom panels consists of high-amplitude ambiguous variables in cluster 2. Since the two algorithms represent mostly the same objects per cluster, we focused our analysis on the clusters from the t-SNE components.

Furthermore, we measured the importance score of each of the 27 features using random forest based on the assigned label for each cluster, as we did with the manually labelled data. The results indicated that the amplitude of the variability in the *G* band (`amp_G`) has the highest feature score, followed by the difference between the highest and lowest values of the *G*-band light curves (`range_mag_g_fov`) and the interquartile range of the *G*-band light curves (`iqr_mag_g_fov`). The rest of the features are listed according to their importance score in Table A.3. As shown in the left panel of Fig. 2, the distribution of the amplitude in a log space reveals three distributions that support these results. In the same figure, a lower bound of the amplitude is shown at ~20 mmag for cluster 0. Additionally, the right panel of Fig. 2 reveals that the amplitude values gradually increase from low to high values of the t-SNE component 1.

Based on these results, we considered all objects in cluster 0 (290 objects) as potential variable hot subdwarfs, and we discuss

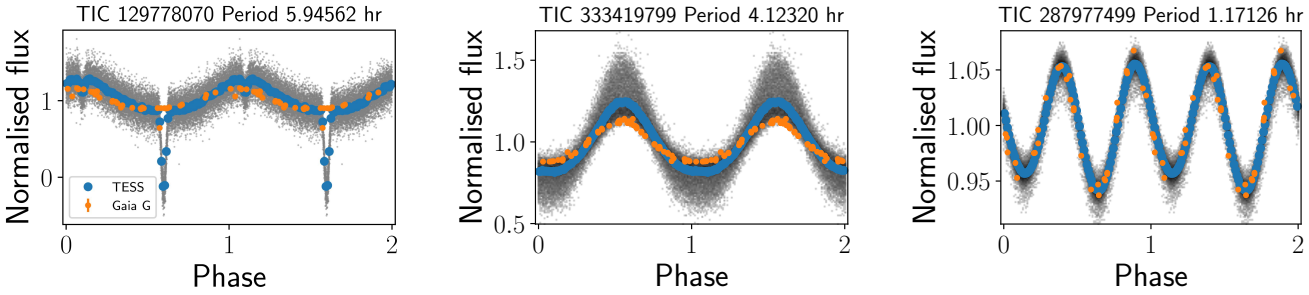


Fig. 3. Three examples of new variable hot subdwarfs identified in this work. The light curves are phase-folded to the same periods and reference epochs. The three objects correspond to an HW Vir (left), a reflection effect (middle), and an ellipsoidal system (right). The light curves on the right are phased to twice the period to highlight the ellipsoidal variation. The blue lines represent the binned phase of the TESS light curves (grey data points), and the orange data points correspond to *Gaia* light curves.

their variability in Sect. 4.1, while objects in cluster 2 were found to be mostly comprised of CVs and are discussed in Sect. 4.2.

4. Results

4.1. Hot subdwarf variability classification

To confirm the nature of the variations found in the *Gaia* light curves, we compared them with those observed by TESS. First, we verified whether any objects in the *Gaia* catalogue had light curves in TESS using the `Lightkurve` Python package (Lightkurve Collaboration 2018). Second, we searched for fast-cadence (20 seconds) and short-cadence (2 minutes) light curves and computed their Lomb-Scargle periodograms.

The periods found in the *Gaia G* band data strongly agree with those obtained by TESS for the objects in cluster 0. The variability types of these objects were thus determined with high confidence. On the other hand, for objects without TESS observations, we are only able to provide a general classification, such as an eclipsing binary or a sinusoidal-like shape class. In order to ensure a homogeneous treatment of the whole sample, we did not rely on TESS data for the results of the frequency analysis. We instead only used the TESS data to improve the fidelity of the classification. All lists of the candidate classifications are provided in Tables A.4–A.10 (see section Data availability).

4.1.1. Variability in the confirmed hot subdwarfs

We found 78 known variable hot subdwarfs amongst the 290 objects in cluster 0 by cross-matching our data with a catalogue of spectroscopically identified hot subdwarfs and known variable hot subdwarfs from the literature (Schaffenroth et al. 2019, 2022, 2023; Culpan et al. 2022; Barlow et al. 2022; Lei et al. 2023; Dawson et al. 2024). Most of them (66/78) were identified from the compiled catalogue of 6616 known hot subdwarfs Culpan et al. (2022), and 63/78 have short- or fast-cadence TESS ligh tcurves. Based on the *Gaia* and TESS light curves, we found 32 reflection-effect systems, 19 HW Vir systems, 6 pulsating variables, and 6 ellipsoidal variables. The remaining 15/78 systems were classified based solely on the *Gaia* three-band light curves, where we found 5 sinusoidal-like light curves that might be associated with reflection-effect systems or ellipsoidal variations or a dominant pulsation mode, 5 eclipsing binaries, and 2 HW Vir systems. Fig. 3 shows examples of new HW Vir (TIC 129778070), reflection effect (TIC 333419799), and ellipsoidal variables (TIC 287977499) systems identified in this work.

4.1.2. Variability in the candidate hot subdwarfs

From the unconfirmed hot subdwarfs (212/290), we identified 78 objects with short- and/or fast-cadence TESS light curves. Based on the *Gaia* and TESS light curves, we found 42 reflection-effect systems, 21 HW Vir systems, 3 pulsating variables, and 2 ellipsoidal variables. The remaining 134/212 candidate hot subdwarfs were classified based on the *Gaia* three-band light curves, where we found 60 sinusoidal-like light curves, 20 HW Vir systems, 14 eclipsing binaries, and 2 potentially pulsating variables. Thirty-eight objects have an unclear variability, which prevented us from labelling them.

4.1.3. Pulsating hot subdwarfs

We identified a total of nine already known pulsating variables from the known and candidate hot subdwarfs observed from *Gaia* and TESS. Three out of these nine pulsate in the *Gaia* and TESS light curves, namely TIC 273218137, TIC 53826859, and TIC 178626010, with a period of 0.09491 h, 0.12096 h, and 1.39841 h, respectively. TIC 273218137 and TIC 53826859 are known pulsating hot subdwarfs from TESS observations (Krzesinski & Balona 2022), while TIC 178626010 is a new pulsating variable detected in this work and independently by Krzesinski et al. (in prep.). In Fig. 4, their *Gaia* and TESS light curves are phased to the same periods and reference epochs t_0, using the short-cadence light curves for TESS observations. The dominant frequencies found for these two objects are the same in the three *Gaia* bands. Therefore, they are candidates for a mode identification from an amplitude ratio analysis (Aerts & Tkachenko 2024; Fritzewski et al. 2024). Their pulsation frequencies suggest that TIC 273218137 and TIC 53826859 are likely *p*-mode pulsators, and TIC 178626010 pulsates in the *g*-mode regime. The remaining six known pulsating variables have low-amplitude pulsations and higher-amplitude orbital variability in their light curves. In our analysis, we were only able to detect their orbital variability in the *Gaia* data.

4.1.4. Newly identified pulsating variables

We identified two unique high-amplitude pulsating objects from *Gaia* (Fig. 5): *Gaia* DR3 5835161264415038592 and *Gaia* DR3 5929109825689001856 with *G*-band peak-to-peak amplitudes of 0.21 mag and 0.25 mag and pulsation periods of 0.38225 h (22.935 min) and 0.12844 h (7.706 min), respectively. The *BP* and *RP* periods for the two objects are the same as those determined in the *G* band. Their amplitudes in these bands are as follows: *Gaia* DR3 5835161264415038592 has peak-to-peak

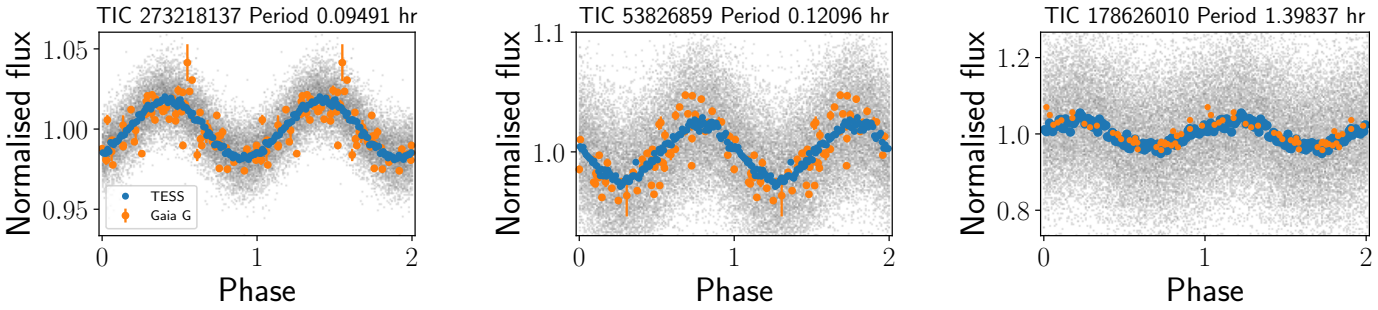


Fig. 4. Three pulsating hot subdwarfs observed with *Gaia* and TESS. The left and middle panels correspond to known pulsating variables (Krzesinski & Balona 2022), and the right panel shows a pulsating variable identified in this work and Krzesinski et al (in prep.). The two light curves are folded to the same periods and reference epochs. The blue lines represent the binned phase of the TESS light curves (grey data points), and the orange data points correspond to *Gaia* light curves.

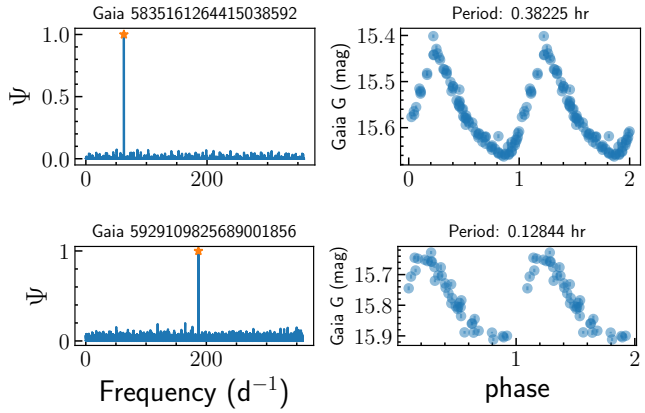


Fig. 5. New high-amplitude pulsating variables observed with *Gaia*.

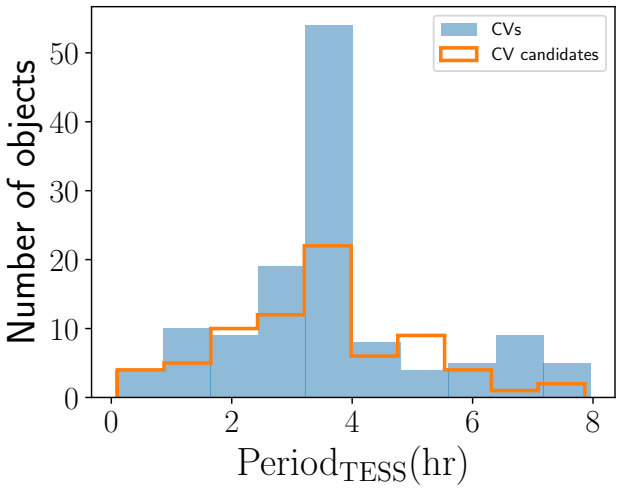


Fig. 6. Period distribution of the confirmed and candidate CVs in cluster 2.

amplitudes of 0.21 mag and 0.19 mag in the *BP* and *RP* bands, respectively. Similarly, *Gaia* DR3 5929109825689001856 has peak-to-peak amplitudes of 0.29 mag and 0.23 mag in the *BP* and *RP* bands, respectively. Their amplitude and frequency regimes suggest that these are candidate blue large amplitude pulsators (BLAPs; Pietrukowicz et al. 2017; Macfarlane et al. 2015).

4.2. Cataclysmic variables

Cluster 2 consists of 296 objects, 140 of which are known CVs (Barlow et al. 2022; Hou et al. 2023; Canbay et al. 2023) and 4 are candidate CVs from Krzesinski et al. (in prep.). The remaining 152 objects are identified by SIMBAD as candidate hot subdwarfs (70), stars (61), variables (9), and CV candidates (3). We considered all of these objects as candidate CVs since all known objects in cluster 2 are CVs without contamination from other classes. The full lists of confirmed and candidate CVs are given in Table A.8 and A.9, respectively.

By cross-matching the objects in cluster 2 with TESS, we found 127/140 confirmed CVs and 75/152 candidate CVs with TESS short-cadence light curves. Their period distributions are shown in Fig. 6, where the periods are centred at 3.43 h and 4.63 h for the known and candidate CVs, respectively. The 127/140 CVs represent the same objects as in the (Canbay et al. 2023) catalogue. However, their reported periods are only available for 71 objects, mainly taken from Ritter & Kolb (2003), with a median period of 3.40 h. This means that we added 56 new candidate orbital periods from our analysis.

4.3. Variability distributions

We investigated the photometric variability of 290 and 296 objects in cluster 0 and cluster 2, respectively. A summary of the variability classification of confirmed and candidate hot subdwarfs is presented in Table 1. In Fig. 7 we present a *Gaia* colour-magnitude diagram of the 1576 candidate hot subdwarf variables (grey circles) with the *Gaia* Catalogue of Nearby Stars in the background (grey data points; Gaia Collaboration 2021b). Classified variables from cluster 0 with TESS light curves are shown in the figure. The light-curve shapes of reflection-effect systems can be explained by the fact that the hot subdwarf irradiates and heats one side of its cooler companion star, causing the cooler star to appear brighter on the side facing the hot subdwarf. As the system orbits, this creates a quasi-sinusoidal variability in the light curves. Depending on the viewing angle, reflection-effect systems can be eclipsing and form the HW Vir systems. On the other hand, compact hot subdwarf binaries, particularly those with white dwarf companions, show ellipsoidal modulation in their light curves due to tidal distortion of the hot subdwarf, resulting in two maxima or two minima in their light curves. Examples of a reflection, HW Vir, and ellipsoidal system are shown in Fig. 3. As previously introduced, the evolutionary stages of these systems can be understood through the lens of a binary evolution channel, notably a common-envelope evolution for short-period systems. However, the exact forma-

Table 1. Variability classifications for known and candidate hot subdwarfs.

141/290 variable hot subdwarf candidates with *Gaia* and TESS lightcurves in cluster 0					
	63 confirmed hot subdwarfs		78 hot subdwarf candidates		
	Confirmed variables	New variables	Confirmed variables	New variables	Total new
Reflection	15	17	2	40	57
HW Vir	13	6	7	14	20
Ellipsoidal	1	5	–	2	7
Pulsating variables	6	–	3	1	1
Others/Unclear	1	1	5	5	–
149/290 variable hot subdwarf candidates with only *Gaia* lightcurves in cluster 0					
	15 confirmed hot subdwarfs		134 hot subdwarf candidates		Total new
Sinusoidal	5		60		65
HW Vir	2		20		22
Eclipsing binary	5		14		19
Pulsating variables	–		2		2
Others/Unclear	3		38		–

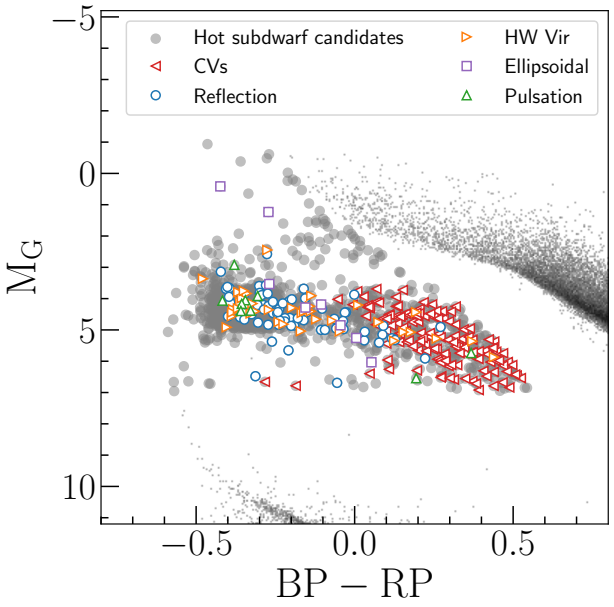


Fig. 7. *Gaia* DR3 colour-magnitude diagram depicting the candidate hot subdwarfs (1682) from Culpan et al. (2022) with *Gaia* light curves (grey circles). The variability classifications are shown for the selected candidate variable hot subdwarfs (290) with TESS observations (141/290). Among the candidate hot subdwarfs, CVs are also identified from the literature and are represented by left triangles. The grey background data points correspond to the *Gaia* Catalogue of Nearby Stars (Gaia Collaboration 2021b).

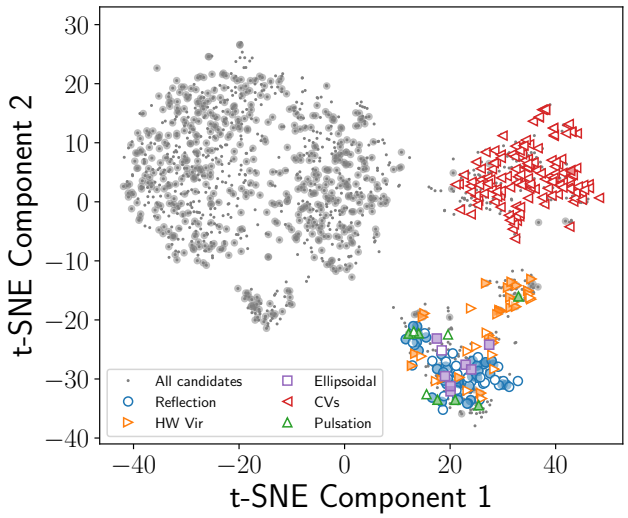


Fig. 8. Identified variables from *Gaia* and TESS light curves. The shaded colours correspond to confirmed hot subdwarfs. CV objects were obtained from the literature (see Sect. 4.2).

tion mechanisms and evolutionary pathways are still areas of active research. On the other hand, CVs consist of a white dwarf primary and a mass-transferring secondary, typically a MS star. The shape of their light curves can mostly be explained by dramatic brightness increases known as outbursts, which are a result of instabilities in the accretion disk and lead to sudden higher mass transfer. In Fig. 7, reflection-effect and HW Vir systems appear to occupy the same area (centred at M_G = 4.4 and $BP{-}RP$ = −0.2) and tend to be bluer than the known CVs (centered at M_G = 5.3 and $BP{-}RP$ = 0.3).

Based on their locations in the t-SNE components, HW Vir systems tend to be more concentrated in the sub-cluster between cluster 0 and cluster 2, as shown in Fig. 8, with a broader *G*-magnitude range (`range_mag_g_fov` around 0.50 mag) compared to the rest of the variables in cluster 0 (`range_mag_g_fov` around 0.16 mag). Poor *Gaia* sampling of HW Vir systems could result in a sinusoidal-like shape of their light curves, as shown in the first panel of Fig. 3, due to the smearing effect. This could place them in a different position in cluster 0 rather than in the sub-cluster. However, some HW Vir systems have shallower eclipse depths compared to others, and this could also place them in the main cluster in cluster 0. As previously mentioned, CVs lie in cluster 2 with a *G*-magnitude range, `range_mag_g_fov`, centred at 1.15 mag. The distributions of the other features are presented in Fig. A.3, with the 10th percentile, the median, and the 90th percentile of the features for each cluster. In comparison to the other two clusters, cluster 2 exhibits a broader distribution of features, notably a high amplitude of variability, as shown in the right panel of Figs. 2 and A.3. These differences in the feature distributions could be relevant for the reduction algorithms to represent the clusters in the low-dimensional space well.

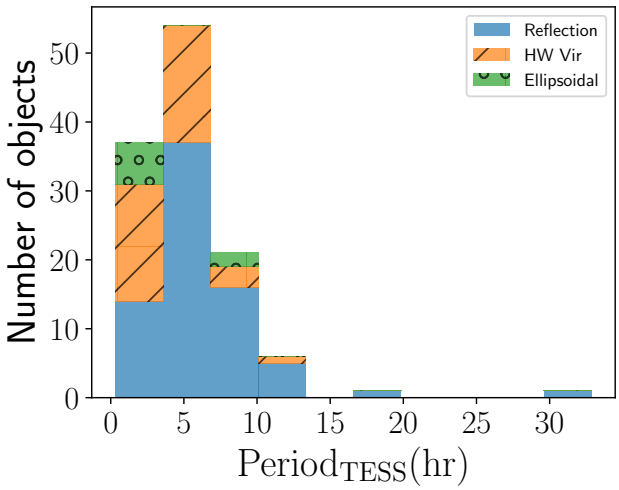


Fig. 9. Period distribution of the binary systems observed with *Gaia* and TESS in cluster 0.

Of the new variables identified from *Gaia* and TESS in cluster 0, ∼23% are classified as HW Vir systems, ∼67% are reflection-effect systems, and ∼10% are ellipsoidal and pulsating variables. For their period distributions, Fig. 9 shows that the periods of known and new HW Vir systems are in the range of ∼1.5 h to ∼9 h, while those of the reflection-effect systems range from ∼1.7 h to ∼35 h. This difference in the period distribution of the eclipsing reflection-effect (HW Vir) and non-eclipsing reflection-effect systems has also been observed in other studies. HW Vir systems tend to have shorter periods than non-eclipsing reflection-effect systems, as found by Schaffenroth et al. (2022). These authors also found a broad peak at periods from 2 to 8 h, but were unable to find objects with a period longer than ∼30 h for reflection-effect systems. They reported that periods longer than a few days might be rare or might not exist for these systems. However, we found several objects with periods longer than a few days from *Gaia*, which might be binary or eclipsing systems. Since we have no TESS observations for these objects, their variability types are referred to as sinusoidal or eclipsing binary.

5. Conclusion and future prospects

We set out to develop a machine-learning algorithm that might be generalised and that leverages multi-band photometric time-series data in order to classify variable and non-variable subdwarfs. We developed our algorithm using multi-band time-series photometry from *Gaia* and validated the algorithm using independent TESS data. Starting with a readily available catalogue of 61 585 candidate hot subdwarfs, we were able to extract *Gaia* multi-band light curves of 1682 objects with good astrometric solutions and a variable number of observations in the *Gaia* photometric bands (with 25 observations at least). We searched for periodicities using the hybrid Ψ-statistic approach and estimated the uncertainties associated with the determined frequencies with a Monte Carlo approach.

Using the sparsely sampled multi-band *Gaia* photometric data, we defined a number of bespoke summary statistics to supplement those already provided by the *Gaia* database. We applied machine-learning algorithms to calculate the importance of the feature and reduce the dimensionality before we applied a clustering algorithm that identified three clusters, which are predominantly predicted by the amplitude of the photometric variability in the *Gaia G* band. We further validated the results by applying two different dimensionality reduction techniques, which resulted in 99% similar results.

The three clusters that we identified correspond to (candidate) hot subdwarfs with statistically significant variability (cluster 0), non-variable subwarfs (cluster 1), and CVs (cluster 2).

Upon further inspection, we were able to identify different populations of variable hot subdwarfs observed from *Gaia* and TESS in cluster 0. A significant number of them are in binaries, while a few pulsating variables are detected. The scarcity of the observed pulsating variables in *Gaia* could be explained by the fact that hot subdwarfs pulsate with low-amplitude light variations of about a few milli-magnitudes.

Our analysis allowed us to newly identify a large number stars as variables, notably reflection-effect and HW Vir systems. The key findings of the clustering analysis are summarised below.

- In cluster 0, 89 new hot subdwarf variables were identified from *Gaia* and TESS observations, while 108 new variables were found from *Gaia* alone. These new variables are mainly reflection-effect and HW Vir systems.
- In the same cluster 0, nine previously identified pulsating variables were found among the candidate variable hot subdwarf. We further identified two new high-amplitude pulsating objects that are consistent with being BLAPs.
- In cluster 2, a large number of CVs were identified, of which 140 were spectroscopically confirmed in other studies. We consider the remaining 156 objects in cluster 2 to be candidate CVs.
- Feature evaluation based on the three clusters showed that features related to the photometric variations in the *G* band strongly contribute to characterising the clusters, including the amplitude, the magnitude range, and the interquartile range of the *G*-band light curves. The *G*-band amplitude distribution suggests a lower limit of ∼0.02 mag on the detection of clear variability in the light curve.

The classification algorithm developed in this work was specifically designed to be flexible and generalisable. We used widely available features and developed new features that can be efficiently calculated for independent data sets with different properties. As a result of this, we can include new observations and objects without having to retrain the algorithm. Furthermore, our results can be used to help build labelled datasets for future supervised machine-learning classifications of variable stars.

Scientifically, our results are twofold. First, we developed a robust method for identifying variable subdwarf stars. Second, we developed an algorithm that efficiently identifies CVs without the need for expensive follow-up spectroscopic observations. Together, these results allowed us to confidently identify new variable subdwarfs for further analysis from existing data while filtering out contaminating sources such as CVs. While hundreds of hot subdwarfs and CVs have already been discovered, a systematically discovered sample of these objects is required to better understand various binary interaction processes, such as mass transfer, common-envelope evolution, and tidal interactions. Furthermore, an algorithm that efficiently identifies variable and non-variable subdwarfs from sparsely sampled data with known amplitude biases offers a unique opportunity for building observational instability strips. By increasing the number of known sdBVs, we can perform population-level asteroseismic studies, similar to the work done for β Cep stars using *Gaia* and TESS data (Fritzewski et al. 2024). This approach has the potential to reveal new insights into the pulsation properties and interior structure of hot subdwarfs by leveraging multi-colour photometry and observational amplitude ratios for mode identifications.

Spectroscopic follow-up observations, such as those with the 4-metre Multi-Object Spectroscopic Telescope (4MOST; de Jong et al. 2019), the William Herschel Telescope Enhanced Area Velocity Explorer (WEAVE; Jin et al. 2024), the Sloan Digital Sky Survey V (SDSS-V; Kollmeier et al. 2019), and the Large sky Area Multi-Object fiber Spectroscopic Telescope (LAMOST; Cui et al. 2012) may deliver radial velocity data and atmospheric parameters to confirm the physical nature of these new variables (153 candidate hot subdwarf and 152 candidate CVs), as well as the two new high-amplitude pulsating variables identified from *Gaia*. Other future prospects include photometric observations of the pulsating variables identified in this work using BlackGEM (Groot et al. 2024) to obtain multi-band pulsation amplitudes for mode identifications and asteroseismic modelling. Additionally, the release of *Gaia* Data Release 4 (DR4), which will include all photometric data, offers a valuable prospect for further exploration. When the complete photometric dataset becomes available, this work can immediately be applied to the remaining 59 471 objects, enabling a comprehensive analysis of variability across a wider range of sources.

Data availability

Tables A.4–A.10 are available at the CDS via anonymous ftp to `cdsarc.cds.unistra.fr` (`130.79.128.5`) or via `https://cdsarc.cds.unistra.fr/viz-bin/cat/J/A+A/693/A268`

Acknowledgements. TK acknowledges support from the National Science Foundation through grant AST #2107982, from NASA through grant 80NSSC22K0338 and from STScI through grant HST-GO-16659.002-A. Co-funded by the European Union (ERC, CompactBINARIES, 101078773). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. The research leading to these results has received funding from the Research Foundation Flanders (FWO) under grant agreement G0A2917N (BlackGEM), as well as from the BELgian federal Science Policy Office (BELSPO) through PRODEX grants for *Gaia* data exploitation. This work has made use of data from the European Space Agency (ESA) mission *Gaia* (`https://www.cosmos.esa.int/Gaia`), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC, `https://www.cosmos.esa.int/web/Gaia/dpac/consortium`). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the *Gaia* Multilateral Agreement. PJG is supported by NRF SARChI grant 111692.

References

Aerts, C., & Tkachenko, A. 2024, A&A, 692, R1
Baluev, R. V. 2008, MNRAS, 385, 1279
Barlow, B. N., Corcoran, K. A., Parker, I. M., et al. 2022, ApJ, 928, 20
Bellm, E. C., Kulkarni, S. R., Graham, M. J., et al. 2019, PASP, 131, 018002
Bloemen, S., Hu, H., Aerts, C., et al. 2014, A&A, 569, A123
Brassard, P., Fontaine, G., Billères, M., et al. 2001, ApJ, 563, 1013
Breiman, L. 2001, Mach. Learn., 45, 5
Canbay, R., Bilir, S., Özdönmez, A., & Ak, T. 2023, AJ, 165, 163
Charpinet, S., Fontaine, G., Brassard, P., et al. 1997, ApJ, 483, L123
Charpinet, S., Green, E. M., Baglin, A., et al. 2010, A&A, 516, L6
Cui, X.-Q., Zhao, Y.-H., Chu, Y.-Q., et al. 2012, RAA, 12, 1197
Cui, K., Liu, J., Feng, F., & Liu, J. 2022, AJ, 163, 23
Culpan, R., Geier, S., Reindl, N., et al. 2022, A&A, 662, A40
Dawson, H., Geier, S., Heber, U., et al. 2024, A&A, 686, A25
Deca, J., Marsh, T. R., Østensen, R. H., et al. 2012, MNRAS, 421, 2798
de Jong, R. S., Agertz, O., Berbel, A. A., et al. 2019, The Messenger, 175, 3
Dorman, B., Rood, R. T., & O'Connell, R. W. 1993, ApJ, 419, 596
Eyer, L., Audard, M., Holl, B., et al. 2023, A&A, 674, A13
Fontaine, G., Brassard, P., Charpinet, S., et al. 2003, ApJ, 597, 518
Fritzewski, D. J., Vanrespaille, M., Aerts, C., Hey, D., & De Ridder, J. 2024, A&A, submitted [arXiv:2408.06097]
Gaia Collaboration (Brown, A. G. A., et al.) 2018, A&A, 616, A1
Gaia Collaboration (Brown, A. G. A., et al.) 2021a, A&A, 649, A1
Gaia Collaboration (Smart, R. L., et al.) 2021b, A&A, 649, A6
Gaia Collaboration (Vallenari, A., et al.) 2023, A&A, 674, A1
Geier, S. 2020, A&A, 635, A193
Geier, S., Heber, U., Tillich, A., et al. 2010, Ap&SS, 329, 91
Geier, S., Raddi, R., Gentile Fusillo, N. P., & Marsh, T. R. 2019, A&A, 621, A38
Geier, S., Dorsch, M., Pelisoli, I., et al. 2022, A&A, 661, A113
Geier, S., Dorsch, M., Dawson, H., et al. 2023, A&A, 677, A11
Ginsburg, A., Sipőcz, B. M., Brasseur, C. E., et al. 2019, AJ, 157, 98
Groot, P. J., Bloemen, S., Vreeswijk, P. M., et al. 2024, PASP, 136, 115003
Han, Z., Podsiadlowski, P., Maxted, P. F. L., Marsh, T. R., & Ivanova, N. 2002, MNRAS, 336, 449
Han, Z., Podsiadlowski, P., Maxted, P. F. L., & Marsh, T. R. 2003, MNRAS, 341, 669
Heber, U. 2009, ARA&A, 47, 211
Heber, U. 2016, PASP, 128, 082001
Hou, W., Luo, A. L., Dong, Y.-Q., Chen, X.-L., & Bai, Z.-R. 2023, AJ, 165, 148
Hu, H., Dupret, M. A., Aerts, C., et al. 2008, A&A, 490, 243
Hu, H., Tout, C. A., Glebbeek, E., & Dupret, M. A. 2011, MNRAS, 418, 195
Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111
Jin, S., Trager, S. C., Dalton, G. B., et al. 2024, MNRAS, 530, 2688
Kao, W.-B., Zhang, Y., & Wu, X.-B. 2024, PASJ, 76, 653
Kim, D.-W., Yeo, D., Bailer-Jones, C. A. L., & Lee, G. 2021, A&A, 653, A22
Kollmeier, J., Anderson, S. F., Blanc, G. A., et al. 2019, BAAS, 51, 274
Krzesinski, J., & Balona, L. A. 2022, A&A, 663, A45
Kuhn, M., & Johnson, K. 2019, Feature Engineering and Selection: A Practical Approach for Predictive Models (Boca Raton: Chapman and Hall/CRC)
Kupfer, T., Geier, S., Heber, U., et al. 2015, A&A, 576, A44
Lei, Z., Zhao, J., Németh, P., & Zhao, G. 2020, ApJ, 889, 117
Lei, Z., He, R., Németh, P., et al. 2023, ApJ, 942, 109
Liao, H., Ren, G., Chen, X., Li, Y., & Li, G. 2024, AJ, 167, 180
Lightkurve Collaboration (Cardoso, J. V. D. M., et al.) 2018, Astrophysics Source Code Library [record ascl:1812.013]
Luo, Y., Németh, P., Deng, L., & Han, Z. 2019, ApJ, 881, 7
Macfarlane, S. A., Toma, R., Ramsay, G., et al. 2015, MNRAS, 454, 507
McInnes, L., Healy, J., & Melville, J. 2018a, ArXiv e-prints [arXiv:1802.03426]
McInnes, L., Healy, J., Saul, N., & Grossberger, L. 2018b, J. Open Source Softw., 3, 861
Monsalves, N., Jaque Arancibia, M., Bayo, A., et al. 2024, A&A, 691, A106
Morales-Rueda, L., Groot, P. J., Augusteijn, T., et al. 2006, MNRAS, 371, 1681
Ostrowski, J., Baran, A. S., Sanjayan, S., & Sahoo, S. K. 2021, MNRAS, 503, 4646
Pantoja, R., Catelan, M., Pichara, K., & Protopapas, P. 2022, MNRAS, 517, 3660
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, J. Mach. Learn. Res., 12, 2825
Pelisoli, I., Vos, J., Geier, S., Schaffenroth, V., & Baran, A. S. 2020, A&A, 642, A180
Pietrukowicz, P., Dziembowski, W. A., Latour, M., et al. 2017, Nat. Astron., 1, 0166
Pojmanski, G. 2002, Acta Astron., 52, 397
Ranaivomanana, P., Johnston, C., Groot, P. J., et al. 2023, A&A, 672, A69
Reed, M. D., Shoaf, K. A., Németh, P., et al. 2020, MNRAS, 493, 5162
Richards, J. W., Starr, D. L., Miller, A. A., et al. 2012, ApJS, 203, 32
Ricker, G. R., Winn, J. N., Vanderspek, R., et al. 2015, J. Astron. Telesc. Instrum. Syst., 1, 014003
Ritter, H., & Kolb, U. 2003, A&A, 404, 301
Rousseeuw, P. J. 1987, J. Comput. Appl. Math., 20, 53
Saffer, R. A., Bergeron, P., Koester, D., & Liebert, J. 1994, ApJ, 432, 351
Saha, A., & Vivas, A. K. 2017, AJ, 154, 231
Sahoo, S. K., Baran, A. S., Heber, U., et al. 2020, MNRAS, 495, 2844
Scargle, J. D. 1982, ApJ, 263, 835
Schaffenroth, V., Barlow, B. N., Geier, S., et al. 2019, A&A, 630, A80
Schaffenroth, V., Pelisoli, I., Barlow, B. N., Geier, S., & Kupfer, T. 2022, A&A, 666, A182
Schaffenroth, V., Barlow, B. N., Pelisoli, I., Geier, S., & Kupfer, T. 2023, A&A, 673, A90
Schwarzenberg-Czerny, A. 1996, ApJ, 460, L107
Silvotti, R., Németh, P., Telting, J. H., et al. 2022, MNRAS, 511, 2201
Steeghs, D., Galloway, D. K., Ackley, K., et al. 2022, MNRAS, 511, 2405
Uzundag, M., Córsico, A. H., Kepler, S. O., et al. 2021, A&A, 655, A27
Uzundag, M., Silvotti, R., Baran, A. S., et al. 2023, Bulletin de la Societe Royale des Sciences de Liege, 92, 11294
Uzundag, M., Krzesinski, J., Pelisoli, I., et al. 2024, A&A, 684, A118
van der Maaten, L., & Hinton, G. 2008, J. Mach. Learn. Res., 9, 2579
VanderPlas, J. T. 2018, ApJS, 236, 16
Van Grootel, V., Charpinet, S., Fontaine, G., Green, E. M., & Brassard, P. 2010, A&A, 524, A63
Vos, J., Vučković, M., Chen, X., et al. 2019, MNRAS, 482, 4592
Vos, J., Bobrick, A., & Vučković, M. 2020, A&A, 641, A163
Webbink, R. F. 1984, ApJ, 277, 355

Appendix A: Additional material

Table A.1. All 84 features used in the feature selection.

No.	Feature	Description
		Selected features for the clustering analysis
1	log_sigvar*	Significance of variability in the G band in a log scale
2	frac_period*	Period over the standard deviation (std) of the three band *Gaia* lightcurve periods
3	std*	Standard deviation of the G, BP, and RP periods
4	fapG*	False alarm probability of the Lomb-Scargle dominant frequency peak (G band)
5	fapRP*	False alarm probability of the Lomb-Scargle dominant frequency peak (RP band)
6	fapBP*	False alarm probability of the Lomb-Scargle dominant frequency peak (BP band)
7	Period_G*	Derived period from the G-band lightcurve
8	Period_BP*	Derived period in the BP-band lightcurve
9	period_RP*	Derived period in the RP-band lightcurve
10	amp_G*	Amplitude of variability in the G band (mag.)
11	amp_BP*	Amplitude of variability in the BP band (mag.)
12	kurtosisG*	G-band kurtosis of the periodogram
13	p99*	99th percentile of all periodogram peaks based on the G-band lightcurves
14	p95_100*	95th percentile of the first 100 frequency peaks based on the G-band lightcurves
15	n05*	Number of peaks above 0.5 of the normalised $\Psi-$ periodogram based on the G band
16	psi_sigvar*	G-band median absolute deviation of the periodogram
17	bp_rp †	BP−RP colour
18	range_mag_g_fov	The range of the G-band time series
19	abbe_mag_g_fov	The Abbe value of the G-band time series
20	iqr_mag_g_fov	The Interquartile Range (IQR) of the G-band time series
21	mad_mag_g_fov	The Median Absolute Deviation (MAD) of the G-band time series
22	stetson_mag_g_fov	The single-band Stetson variability index
23	abbe_mag_bp	The Abbe value of the BP-band time series
24	abbe_mag_rp	The Abbe value of the RP-band time series
25	outlier_median_g_fov	Greatest absolute deviation from the G median normalised by the error
26	skewness_mag_bp	The standardised unbiased unweighted skewness of the BP-band time series
27	std_dev_over_rms_err_mag_g_fov	S/N ratio G estimate
		Excluded features in the feature selection processes
28	G_abs*	Gaia G absolute magnitude
29	N_G*	Number of observations in the G band.
30	N_BP*	Number of observations in the BP band
31	N_RP*	Number of observations in the RP band.
32	amp_RP*	Amplitude of variability in the RP band (mag.)
33	p90_100*	90th percentile of the first 100 frequency peaks
34	p99_100*	99th percentile of the first 100 frequency peaks
35	rmse_over_ptp_amp*	Root mean square error (RMSE) of the Lomb-Scargle model fit over the peak-to-peak G amplitude
36	parallax †	Gaia parallax
37	parallax_error †	Gaia parallax error
38	phot_g_mean_mag †	G-band mean magnitude
39	phot_g_n_obs †	Number of observation contributing to G photometry
40	RUWE †	Renormalised unit weight error
41	num_selected_g_fov	Total number of G FOV transits selected for variability analysis
42	mean_obs_time_g_fov	Mean observation time for G observations
43	time_duration_g_fov	Time duration of the G time series
44	min_mag_g_fov	The minimum value of the G-band time series
45	max_mag_g_fov	The maximum value of the G-band time series
46	mean_mag_g_fov	The mean of the G-band time series
47	median_mag_g_fov	The median of the G-band time series
48	trimmed_range_mag_g_fov	Trimmed difference between the highest and lowest G-band time series
49	std_dev_mag_g_fov	Square root of the unweighted G magnitude variance
50	skewness_mag_g_fov	The standardised unbiased unweighted skewness of the G-band time series
51	kurtosis_mag_g_fov	The standardised unbiased unweighted kurtosis of the G-band time series
52	num_selected_bp	Total number of BP observations selected for variability analysis
53	mean_obs_time_bp	Mean observation time for BP observations
54	time_duration_bp	Time duration of the BP time series
55	min_mag_bp	The minimum value of the BP-band time series
56	max_mag_bp	The maximum value of the BP-band time series
57	mean_mag_bp	The mean of the BP-band time series
58	median_mag_bp	The median of the BP-band time series
59	range_mag_bp	The range of the BP-band time series
60	trimmed_range_mag_bp	Trimmed difference between the highest and lowest BP-band time series
61	std_dev_mag_bp	Square root of the unweighted BP magnitude variance
62	kurtosis_mag_bp	The standardised unbiased unweighted kurtosis of the BP-band time series
63	mad_mag_bp	The Median Absolute Deviation (MAD) of the BP-band time series
64	iqr_mag_bp	The Interquartile Range (IQR) of the BP-band time series
65	stetson_mag_bp	The single-band Stetson variability index
66	std_dev_over_rms_err_mag_bp	S/N ratio BP estimate
67	outlier_median_bp	Greatest absolute deviation from the BP median normalised by the error

Table A.1. continued.

No.	Feature	Description
68	num_selected_rp	Total number of RP observations selected for variability analysis
69	mean_obs_time_rp	Mean observation time for RP observations
70	time_duration_rp	Time duration of the RP time series
71	min_mag_rp	The minimum value of the RP-band time series
72	max_mag_rp	The maximum value of the RP-band time series
73	mean_mag_rp	The mean of the RP-band time series
74	median_mag_rp	The median of the RP-band time series
75	range_mag_rp	The range of the RP-band time series
76	trimmed_range_mag_rp	Trimmed difference between the highest and lowest RP-band time series
77	std_dev_mag_rp	Square root of the unweighted RP magnitude variance
78	skewness_mag_rp	The standardised unbiased unweighted skewness of the RP-band time series
79	kurtosis_mag_rp	The standardised unbiased unweighted kurtosis of the RP-band time series
80	mad_mag_rp	The Median Absolute Deviation (MAD) of the RP-band time series
81	iqr_mag_rp	The Interquartile Range (IQR) of the RP-band time series
82	stetson_mag_rp	The single-band Stetson variability index
83	std_dev_over_rms_err_mag_rp	S/N ratio RP estimate
84	outlier_median_rp	Greatest absolute deviation from the RP median normalised by the error

Notes. Features marked with (*) were computed in this work, those with (†) are from the Gaia DR3 source database (Gaia Collaboration 2023), while the rest were obtained from the *Gaia* variability summary table (Eyer et al. 2023). A full description of these *Gaia* statistics can be found in the *Gaia* documentation here.

Table A.2. Feature ranking based on the manual labelling. (Eyer et al. 2023)

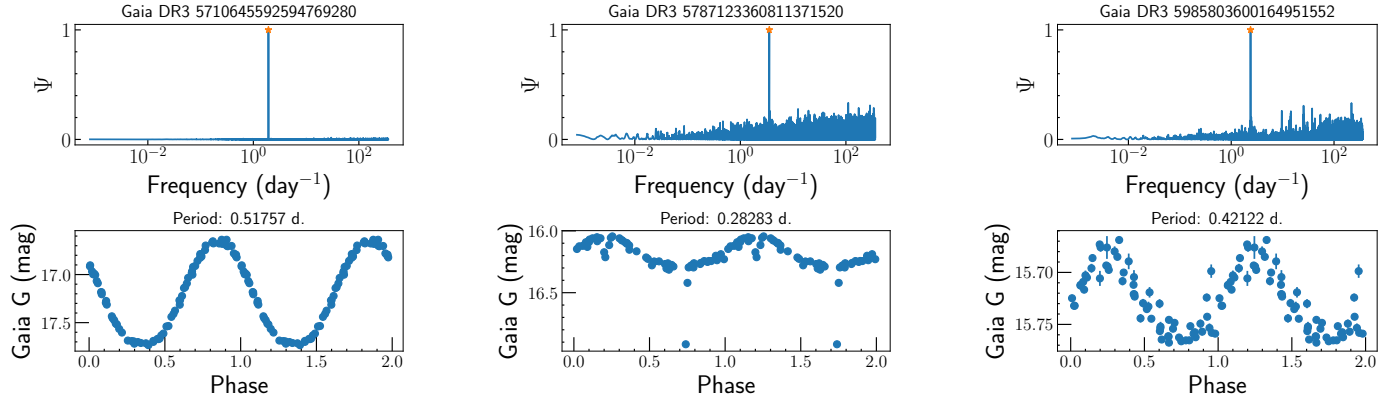
ID	Feature	Description
1	p95_100*	95th percentile of the first 100 frequency peaks based on the G-band lightcurves
2	n05*	Number of peaks above 0.5 of the normalised Ψ periodogram based on the G band
3	p99*	99th percentile of all periodogram peaks based on the G-band lightcurves
4	Period_G*	Derived period from the G-band lightcurve
5	frac_period*	Period over the standard deviation (std) of the three band Gaia lightcurve periods
6	fapG*	False alarm probability of the Lomb-Scargle dominant frequency peak (G band)
7	psi_sigvar*	G-band median absolute deviation of the periodogram
8	kurtosisG*	G-band kurtosis of the periodogram
9	iqr_mag_g_fov	The Interquartile Range (IQR) of the G-band time series
10	std*	Standard deviation of the G, BP, and RP periods
11	amp_G*	Amplitude of variability in the G band (mag.)
12	log_sigvar*	Significance of variability in the G band in a log scale
13	mad_mag_g_fov	The Median Absolute Deviation (MAD) of the G-band time series
14	range_mag_g_fov	The range of the G-band time series
15	abbe_mag_bp	The Abbe value of the BP-band time series
16	abbe_mag_rp	The Abbe value of the RP-band time series
17	Period_RP*	Derived period from the RP-band lightcurve
18	fapBP*	False alarm probability of the Lomb-Scargle dominant frequency peak (BP band)
19	abbe_mag_g_fov	The Abbe value of the G-band time series
20	stetson_mag_g_fov	Stetson G FoV variability index
21	Period_BP*	Derived period from the BP-band lightcurve
22	amp_BP*	Amplitude of variability in the BP band(mag.)
23	fapRP*	False alarm probability of the Lomb-Scargle dominant frequency peak (RP band)
24	std_dev_over_rms_err_mag_g_fov	S/N ratio G FoV estimate
25	bp_rp †	BP − RP colour
26	outlier_median_g_fov	The most outlying measurement with respect to the median
27	skewness_mag_bp	The standardised unbiased unweighted skewness of the BP-band time series

Notes. Features marked with (*) were computed in this work, those with (†) are from the *Gaia* DR3 source database (Gaia Collaboration 2023), while the rest were obtained from the *Gaia* variability summary table.

Table A.3. Feature ranking based on the three cluster labels.

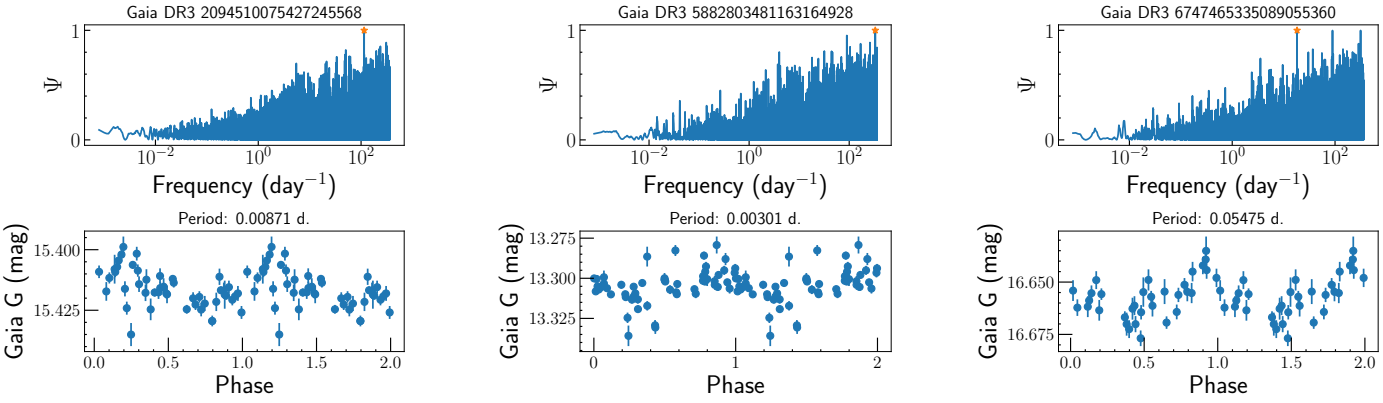
ID	Feature	Description
1	amp_G*	Amplitude of variability in the G band (mag.)
2	range_mag_g_fov	The range of the G-band time series
3	iqr_mag_g_fov	The Interquartile Range (IQR) of the G-band time series
4	log_sigvar*	Significance of variability in the G band in a log scale
5	stetson_mag_g_fov	Stetson G FoV variability index
6	n05*	Number of peaks above 0.5 of the normalised Ψ periodogram based on the G band
7	std_dev_over_rms_err_mag_g_fov	S/N ratio G FoV estimate
8	p95_100*	95th percentile of the first 100 frequency peaks based on the G-band lightcurves
9	mad_mag_g_fov	The Median Absolute Deviation (MAD) of the G-band time series
10	bp_rp ↾	BP − RP colour
11	outlier_median_g_fov	The most outlying measurement with respect to the median
12	p99*	99th percentile of all periodogram peaks based on the G-band lightcurves
13	amp_BP*	Amplitude of variability in the BP band(mag.)
14	Period_G*	Derived period from the G-band lightcurve
15	abbe_mag_g_fov	The Abbe value of the G-band time series
16	kurtosisG*	G-band kurtosis of the periodogram
17	skewness_mag_bp	The standardised unbiased unweighted skewness of the BP-band time series
18	abbe_mag_bp	The Abbe value of the BP-band time series
19	psi_sigvar*	G-band median absolute deviation of the periodogram
20	abbe_mag_rp	The Abbe value of the RP-band time series
21	fapG*	False alarm probability of the Lomb-Scargle dominant frequency peak (G band)
22	Period_RP*	Derived period from the RP-band lightcurve
23	Period_BP*	Derived period from the BP-band lightcurve
24	frac_period*	Period over the standard deviation (std) of the three band Gaia lightcurve periods
25	std*	Standard deviation of the G, BP, and RP periods
26	fapRP*	False alarm probability of the Lomb-Scargle dominant frequency peak (RP band)
27	fapBP*	False alarm probability of the Lomb-Scargle dominant frequency peak (BP band)

Notes. Features marked with (*) were computed in this work, those with (↾) are from the *Gaia* DR3 source database (Gaia Collaboration 2023), while the rest were obtained from the *Gaia* variability summary table (Eyer et al. 2023).



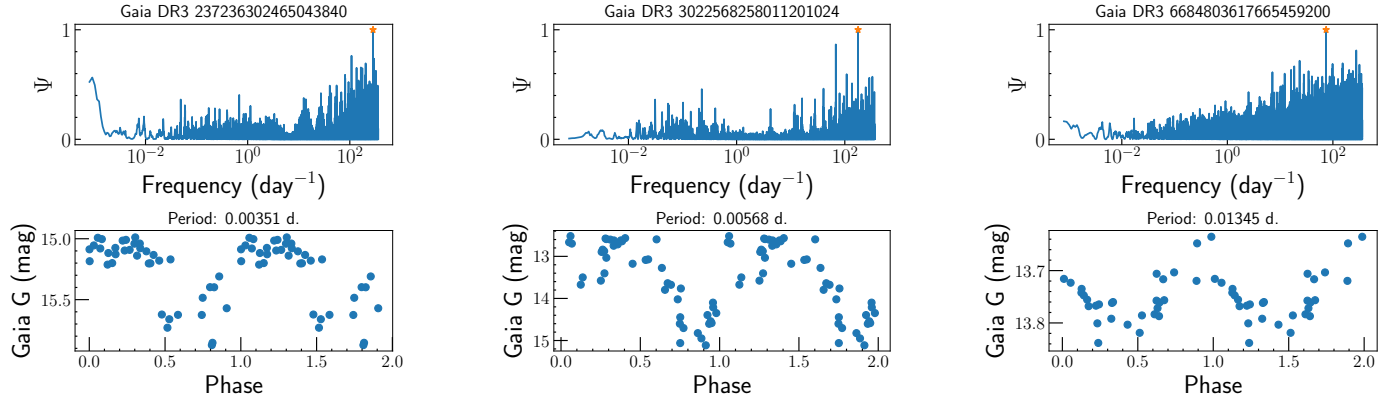
(a) (b) (c)

Three examples of objects (a, b, and c) in cluster 0.



(d) (e) (f)

Three examples of objects (d, e, and f) in cluster 1.



(g) (h) (i)

Three examples of objects (g, h, and i) in cluster 2.

Fig. A.1. Examples of periodograms and phase-folded light curves for each cluster. The top, middle, and bottom rows correspond to cluster 0, cluster 1, and cluster 2, respectively.

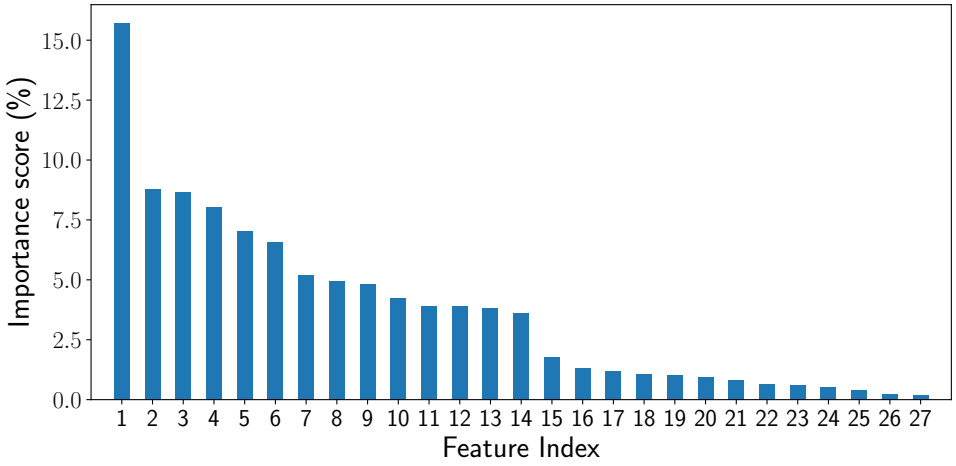


Fig. A.2. Random forest feature importance scores for the 27 features listed in Table A.3. The x-axis corresponds to the Feature ID in the table.

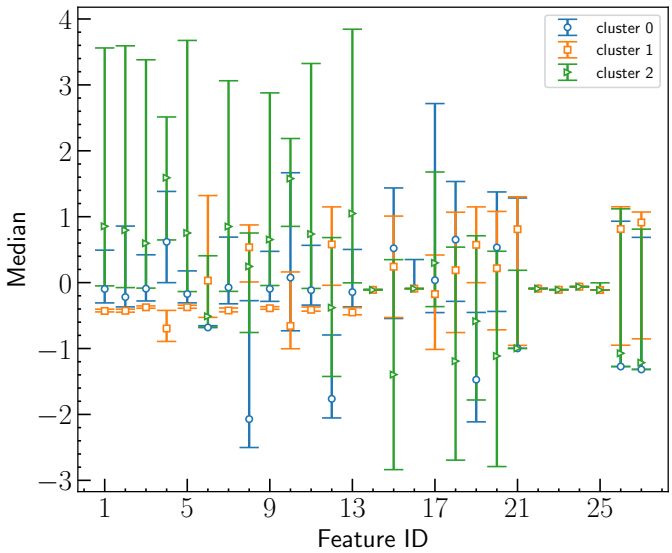


Fig. A.3. Distribution of the feature medians for each cluster. The x-axis corresponds to the feature ID listed in Table A.3; The y-axis represents the median (open marker), the 10th percentile (lower cap), and the 90th percentile (upper cap) of each feature after a z-score normalisation.