Searching for the Weirdest Stars: A Convolutional Autoencoder-Based Pipeline For Detecting Anomalous Periodic Variable Stars

Anonymous Author(s)
Affiliation
Address
email

Abstract

The physical processes of stars are encoded in their periodic pulsations. Millions of variable stars will be observed by the upcoming Vera Rubin Observatory. Here, we present a convolutional autoencoder-based pipeline as an automatic approach to search for anomalous periodic variables within the Zwicky Transient Facility catalog of periodic variable stars. We encode their light curves using the convolutional autoencoder, and we use an isolation forest to score each periodic variable star in the latent space. Our overall most anomalous events share some similarities: they are mostly cool, red, high variability, and irregularly oscillating periodic variables. Observational data suggested that they are most likely young and massive Red Giant or Asymptotic Giant Branch stars. Furthermore, we use the learned latent feature for the classification of periodic variables through a hierarchical random forest. This novel semi-supervised approach allows astronomers to identify the most anomalous events within a given physical class, accelerating the potential for scientific discovery.

1 Introduction

Anomaly detection is a vital aspect for making discoveries in astronomy. Examples include the anomalies in the CMB temperature anisotropies [1], quasi-stellar objects [2] and dark energy [3]. As deep-sky surveys via modern telescopes continue to exponentially increase our discovery rates of galactic and extra-galactic transients, researchers are turning towards automated methods of anomaly detection [4–6].

Advanced techniques to search for anomalous astrophysical events is essential in the era of upcoming observatories. In particular, the Legacy Survey of Space and Time (LSST) conducted by the Vera Rubin Observatory is expected to commence operations in 2024 [7] and is anticipated to observe 32 trillion stars within its 10 years of operation [8]. It is reasonable to expect *anomalous* periodic variables stars (PVSs) which defy expectations. Indeed, their discoveries have already been challenging our understanding of the Galactic metallicity [9], the physics of accretion and mass transfer [10, 11], etc. Despite the impact brought by these discoveries, studies on PVSs that utilize machine learning have only been made on classifications [12–14], and deep generative modeling for parameter estimation [15]. While Malanchev *et al.* [16] searched for anomalous transient detected with the Zwicky Transient Facility (ZTF), they are not specifically aiming for PVSs. Here, we provide an anomaly detection pipeline to effectively search for peculiar PVSs detected with the Zwicky Transient Facility (ZTF). The ZTF collaboration contains numerous publicly available data which serve as a good training set for big data problems in astronomy which will arise with LSST. We anticipated that our pipeline could be applied once the observational data from the Vera Rubin Observatory (or other, new facilities) are made available.

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

2 Method

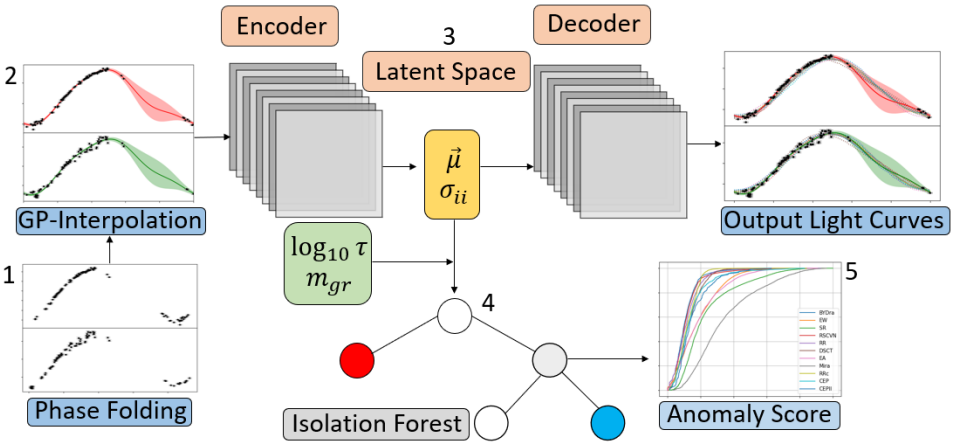


Figure 1: The anomaly detection pipeline: 1. Phase-folding the raw detection data. 2. Interpolation using the MGP. 3. Encoding to get latent vectors $\vec{\mu}$ and matrix elements σ_{ii}. 4. Append $\log_{10}\tau$ and $m_{\rm gr}$ to $\vec{\mu}$. 5. Run the isolation forest and rank the anomalies.

Here we describe our training set and methodology. Our training set consists of the ZTF Variable Star Catalog (ZTF VSC) presented in Chen *et al.* [17]. The ZTF VSC utilizes the Data Release 2 archive of the ZTF [18] to search for and classify new PVSs down to a r-band magnitude, a measurement of brightness, of ~ 20.6. They find a total of $781,602$ PVSs, of which $621,702$ are newly discovered. The data are given in two filters (g-band and r-band) which are observed asynchronously. The ZTF VSC provides periods, which we use to phase-fold the light curves. To phase-fold a light curve, we cut the time series into multiple sub-series with a time duration equal to the period and stack these sub-series on top of each other. Because the data are irregularly sampled by the telescope and taken in both bands, we choose to interpolate the phase-folded light curves using a multivariate Gaussian process (MGP; Foreman-Mackey *et al.* 19, Pruzhinskaya *et al.* 5, Qu *et al.* 20) with periodic boundary conditions. The MGP has a mean function $\eta(\phi, \lambda)$ and a covariance function K. We set $\eta(\phi, \lambda) = 0$, where ϕ is the temporal phase and λ is a wavelength in scaled units. We choose the following covariance function:

$$K(\vec{r}, \vec{r}') = C\exp(-\frac{|\vec{\phi} - \vec{\phi}'|^2}{l_\phi^2})\exp(-\frac{|\vec{\lambda} - \vec{\lambda}'|^2}{l_\lambda^2}) + \begin{cases} \delta & \text{if } \vec{r} = \vec{r}', \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Here, l_ϕ and l_λ measure the correlation along the phase and wavelength direction respectively, C is a constant, and δ measures the white noise level of the raw data. We generate 160 evenly spaced data points along with the phase direction for both g- and r-bands. Following the approach of Villar *et al.* [21] and Villar *et al.* [6], we stack both of them horizontally to form an 'image' of size 2×160. The 'images' will be encoded through the convolutional variational autoencoder [C-VAE, 22] with a LeNet structure[23]. The architecture of the encoder is described as follows:

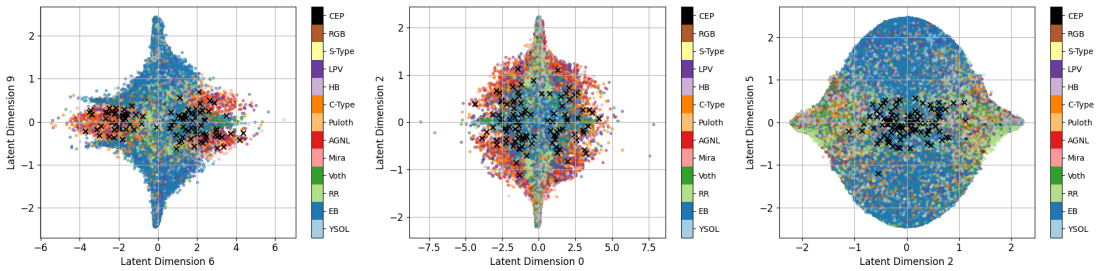
1. **Input layer** of size 2×160
2. **3 Convolutional layers** with the *ReLu activation* with *Dropout*
3. **Dense layer** with 256 neurons, *Linear Activation*
4. **Latent space** of size 2×10

The filter size of the convolutional layer increases from $32 \rightarrow 64 \rightarrow 128$, and the dropout fraction is set to 0.1. Given the variational nature of the autoencoder, the bottle neck latent space is described by a mean vector $\vec{\mu}$ and the diagonal covariance matrix σ_{ii}. Inspired by the work of Zhang and Bloom [13], we apply periodic padding to the 'images' during the convolution to enforce periodic boundary conditions. Our decoder is the symmetric counterpart of the encoder; however, no periodic padding and no dropout is applied. We train with early stopping with a train to validation to test ratio of $7:2:1$. We train for roughly 6 hours on the NVIDIA RTX2080-Ti GPU.

Once trained, we next use an isolation forest to rank each variable by its anomaly score. The isolation forest works by building an ensemble of binary trees, which work to isolate samples of the population. The anomalies are identified as requiring few trees to isolate the event [24]; in this work, we focus on events with the top 100 most anomalous scores. We use the latent vectors $\vec{\mu}$, the log of period $\log_{10}\tau$, and the difference between the average g- and r-band magnitude ($m_{\rm gr} := \langle m_{\rm g}\rangle - \langle m_{\rm r}\rangle$) as input features of the isolation forest [25]. We note that the period and magnitudes are explicitly included because this information is lost in the pre-processing used to train our autoencoder; however, we believe that they will be valuable in filtering out anomalies. We use the isolation forest implemented in the `scikit-learn` package with based estimators of $100,000$. We encourage readers to refer to Figure 1 as a review of our pipeline.

We additionally use our learned latent space to *classify* the ZTF VSC by their physical origin. We note that the ZTF VSC provides class labels for PVSs, but the classifications are based on hand-engineered features. Here we provide an alternative classification method using a hierarchical random forest classifier. We extracted events labels from the SIMBAD catalogue [26] by cross-matching (using the python package `Astroquery` [27]) their sky-coordinates with those listed in the ZTF VSC. We find $31,541$ of successfully cross-matched object. We then construct 13 classes in 2 levels. The first level includes Active Galactic Nuclei-like (AGNL), Cepheid (CEP), Eclipsing Binaries (EB), Long-Period Variables (LPV), Mira variables (Mira), Pulsating Variables (Pul), RR Lyrae (RR) and Peculiar (Pec). The second level are further classifications of the Pec type, and it includes Carbon stars (C-Type), Horizontal Branch stars (HB), Red Giant Branch stars (RGB), S-Type stars (S-Type), Young Stellar Object-like (YSOL) and Variable Others (Voth). We note that our training set is highly imbalanced, with the largest set contain $10,745$ events and the smallest containing just 41 events. The extracted features and the latent vectors $\vec{\mu}$ of the cross-matched objects were fed into the hierarchical random forest provided by the python package `imbalanced-learn` [28], which performs synthetic minority resampling [29, 28]. Finally, the train to test ratio is set to be $7:3$.

3 Results and Discussion

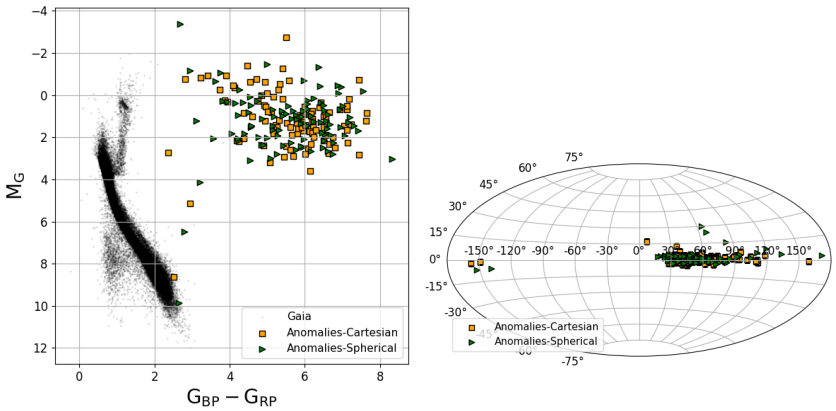


(a) (b) (c)

Figure 2: (a), (b), and (c) Examples of the latent distributions for different PVSs labelled by distinct colours. Anomalies are marked as dark-crosses.

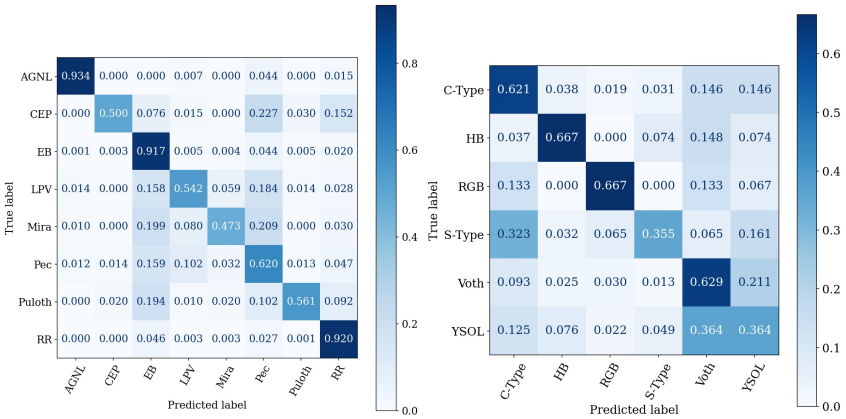
We find that our latent space exhibit an annular structure, which inspires us to transform the latent vector into N-dimensional spherical coordinate. We run our isolation forest on both sets of coordinates and compare results, selecting the top 100 most anomalous events in both cases. Our anomaly detection algorithm is seemingly sensitive to the irregular oscillating which consists of several dominant Fourier modes. In general, these anomalous events are both multi-modal and highly irregular, with some examples exhibiting larger fluctuations that span over several magnitudes. Some of the anomalies also show weak or even anti-correlation between the $g-$ and $r-$ band light curves, suggesting significant temperature variations within the pulsations.

To better understand the nature of the selected anomalies, we extracted Gaia [32] G-band absolute magnitudes $M_{\rm G}$, and the difference between the Gaia B-band and R-band absolute magnitudes $G_{\rm BP} - G_{\rm RP}$ for the top 100 anomalies in both latent spaces. We plot their distribution in Figure 3 (a). We note that this is a common diagnostic phase space, which roughly correlates with temperature and luminosity of the stars. The majority of the anomalies are *cool*, with $G_{\rm BP} - G_{\rm RP} > 4$ and *luminous* with $M_{\rm G} \approx 1$. These properties correspond to seemingly massive ($\simeq 5 - 10{\rm M}_{\odot}$) luminous and cold



(a) (b)

Figure 3: (a) Distribution of the top 100 anomalous events on the Gaia HR diagram (a common feature space in astronomy) against the main-sequence stars in black dots. Data are taken from Lindegren *et al.* [30], Bailer-Jones *et al.* [31], and Gaia Collaboration *et al.* [32], through the Vizier Catalogue [33] by `Astroquery` [27]. (b) Same as (a), but for their distributions in the Galactic coordinates.



(a) (b)

Figure 4: (a) Confusion matrix for the first level classification labels of our test set. (b) Same as (a), but for the second level classification labels.

stars. Furthermore, we show the distribution of the top anomalies in the Galactic coordinates in Figure 3 (b). The majority of the anomalies are lying close to the Galactic plane, implying that (1) there is significant interstellar reddening due to the dust for these events and (2) the progenitor systems of these events are likely young and massive stars. Taken together, the observational evidence points to highly anomalous, young/cool Red Giant or Asymptotic Giant Branch stars. Spectroscopic follow-up observations and detailed light curve modelling is *essential* in fully understanding the anomalies detected in our data-driven pipeline; however, we note that these anomalies were discovered with *limited* survey observations. Similar techniques will be invaluable for future missions.

Finally, we highlight the classification results from our hierarchical classifier in Figure 4 (a) and (b). We find similar results for the baseline comparison with the results presented in the ZTF VSC; however, we classify for substantially more classes with our newly developed method. For example, we find that 97.9% of the AGNL objects (objects associated with supermassive black holes in other galaxies) are labelled as semi-regular *galactic* variable stars in previous studies. In short, we find that our learned latent space is sufficient to classify the set of ZTF PVS light curves explored here. A more detailed diagnosis could help to improve the purity of the ZTF VSC. We plot the latent distribution

using our new labels in Figure 2, in order to show the robustness of our autoencoder in differentiating objects that belongs to different categories.

4 Conclusion

We present a convolutional autoencoder-based pipeline as an automatic yet robust approach to: (1) search for anomalous PVSs within a sea of data. (2) make better classifications for PVS. (3) help improve the purity of the ZTF VSC. Our pipeline can be applied to PVS data obtained from deep sky surveys in the future. Last but not least, our pipeline generates a list of anomalous PVSs. Detailed spectroscopic follow-up is essential to reveal their true identity and how they fit into our current understanding of late-stage stellar processes.

Astrophysical science has entered the big-data era. We anticipate our method can contribute to the community, such as detecting other kinds of anomalous periodic or non-periodic transient, or better classifying different kinds of astronomical objects. Besides, our unsupervised learning pipeline can make use of supervised object labels to look for categorical-wise anomalies, which help astronomers better understand the diversities within each class of objects, and it can improve the labeling and classifications of stellar objects. Techniques similar to those presented here can be used in broader applications in order to identify anomalies in periodic, multi-variate time series.

References

[1] D. C. Rodrigues, Phys. Rev. D **77**, 023534 (2008).

[2] P. Massey, K. F. Neugent, and E. M. Levesque, The Astronomical Journal **157**, 227 (2019).

[3] A. G. Riess, A. V. Filippenko, P. Challis, A. Clocchiatti, A. Diercks, P. M. Garnavich, R. L. Gilliland, C. J. Hogan, S. Jha, R. P. Kirshner, B. Leibundgut, M. M. Phillips, D. Reiss, B. P. Schmidt, R. A. Schommer, R. C. Smith, J. Spyromilio, C. Stubbs, N. B. Suntzeff, and J. Tonry, AJ **116**, 1009 (1998), arXiv:astro-ph/9805201 [astro-ph] .

[4] M. Henrion, D. J. Mortlock, D. J. Hand, and A. Gandy, Classification and anomaly detection for astronomical survey data, in *Astrostatistical Challenges for the New Astronomy*, edited by J. M. Hilbe (Springer New York, New York, NY, 2013) pp. 149–184.

[5] M. V. Pruzhinskaya, K. L. Malanchev, M. V. Kornilov, E. E. O. Ishida, F. Mondon, A. A. Volnova, and V. S. Korolev, Monthly Notices of the Royal Astronomical Society **489**, 3591 (2019), https://academic.oup.com/mnras/article-pdf/489/3/3591/30029513/stz2362.pdf .

[6] V. A. Villar, M. Cranmer, E. Berger, G. Contardo, S. Ho, G. Hosseinzadeh, and J. Yao-Yu Lin, arXiv e-prints , arXiv:2103.12102 (2021), arXiv:2103.12102 [astro-ph.HE] .

[7] M. Graham, in *The Extragalactic Explosive Universe: the New Era of Transient Surveys and Data-Driven Discovery* (2019) p. 23.

[8] Ž. Ivezić *et al.*, The Astrophysical Journal **873**, 111 (2019).

[9] M. I. Jurkovic, Serbian Astronomical Journal **197**, 13 (2018), arXiv:1904.08815 [astro-ph.SR] .

[10] A. Gautschy and H. Saio, Monthly Notices of the Royal Astronomical Society **468**, 4419 (2017), https://academic.oup.com/mnras/article-pdf/468/4/4419/14077699/stx811.pdf .

[11] M. Kimura, S. Yamada, N. Nakaniwa, Y. Makita, H. Negoro, M. Shidatsu, T. Kato, T. Enoto, K. Isogai, T. Mihara, H. Akazawa, K. C. Gendreau, F.-J. Hambsch, P. A. Dubovsky, I. Kudzej, K. Kasai, T. Tordai, E. Pavlenko, A. A. Sosnovskij, J. V. Babina, O. I. Antonyuk, H. Itoh, and H. Maehara, arXiv e-prints , arXiv:2106.15756 (2021), arXiv:2106.15756 [astro-ph.SR] .

[12] S. Jamal and J. S. Bloom, The Astrophysical Journal Supplement Series **250**, 30 (2020).

[13] K. Zhang and J. S. Bloom, Monthly Notices of the Royal Astronomical Society **505**, 515 (2021), https://academic.oup.com/mnras/article-pdf/505/1/515/38334334/stab1248.pdf .

[14] B. Naul, J. S. Bloom, F. Pérez, and S. van der Walt, Nature Astronomy **2**, 151 (2018), arXiv:1711.10609 [astro-ph.IM] .

[15] J. Martínez-Palomera, J. S. Bloom, and E. S. Abrahams, arXiv e-prints , arXiv:2005.07773 (2020), arXiv:2005.07773 [astro-ph.IM] .

[16] K. L. Malanchev, M. V. Pruzhinskaya, V. S. Korolev, P. D. Aleo, M. V. Kornilov, E. E. O. Ishida, V. V. Krushinsky, F. Mondon, S. Sreejith, A. A. Volnova, A. A. Belinski, A. V. Dodin, A. M. Tatarnikov, S. G. Zheltoukhov, and (The SNAD Team), MNRAS **502**, 5147 (2021), arXiv:2012.01419 [astro-ph.IM] .

[17] X. Chen, S. Wang, L. Deng, R. de Grijs, M. Yang, and H. Tian, The Astrophysical Journal Supplement Series **249**, 18 (2020).

[18] E. C. Bellm *et al.*, PASP **131**, 018002 (2019), arXiv:1902.01932 [astro-ph.IM] .

[19] D. Foreman-Mackey, S. Hoyer, J. Bernhard, and R. Angus, george: George (v0.2.0) (2014).

[20] H. Qu, M. Sako, A. Möller, and C. Doux, arXiv e-prints , arXiv:2106.04370 (2021), arXiv:2106.04370 [astro-ph.IM] .

[21] V. A. Villar, G. Hosseinzadeh, E. Berger, M. Ntampaka, D. O. Jones, P. Challis, R. Chornock, M. R. Drout, R. J. Foley, R. P. Kirshner, R. Lunnan, R. Margutti, D. Milisavljevic, N. Sanders, Y.-C. Pan, A. Rest, D. M. Scolnic, E. Magnier, N. Metcalfe, R. Wainscoat, and C. Waters, The Astrophysical Journal **905**, 94 (2020).

[22] D. P. Kingma and M. Welling, arXiv e-prints , arXiv:1312.6114 (2013), arXiv:1312.6114 [stat.ML] .

[23] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, Neural Computation **1**, 541 (1989), https://direct.mit.edu/neco/article-pdf/1/4/541/811941/neco.1989.1.4.541.pdf .

[24] F. T. Liu, K. M. Ting, and Z.-H. Zhou, in *2008 Eighth IEEE International Conference on Data Mining* (2008) pp. 413–422.

[25] F. T. Liu, K. M. Ting, and Z.-H. Zhou, ACM Trans. Knowl. Discov. Data **6**, 10.1145/2133360.2133363 (2012).

[26] M. Wenger, F. Ochsenbein, D. Egret, P. Dubois, F. Bonnarel, S. Borde, F. Genova, G. Jasniewicz, S. Laloë, S. Lesteven, and R. Monier, A&AS **143**, 9 (2000), arXiv:astro-ph/0002110 [astro-ph] .

[27] A. Ginsburg *et al.*, AJ **157**, 98 (2019), arXiv:1901.04520 [astro-ph.IM] .

[28] G. Lemaître, F. Nogueira, and C. K. Aridas, Journal of Machine Learning Research **18**, 1 (2017).

[29] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, Journal of artificial intelligence research **16**, 321 (2002).

[30] L. Lindegren *et al.*, A&A **616**, A2 (2018), arXiv:1804.09366 [astro-ph.IM] .

[31] C. A. L. Bailer-Jones, J. Rybizki, M. Fouesneau, M. Demleitner, and R. Andrae, AJ **161**, 147 (2021), arXiv:2012.05220 [astro-ph.SR] .

[32] Gaia Collaboration *et al.*, A&A **649**, A1 (2021), arXiv:2012.01533 [astro-ph.GA] .

[33] B. A. Skiff, VizieR Online Data Catalog , B/mk (2014).