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

Ho-Sang, Chan
Department of Physics
The Chinese University of Hong Kong
Sha Tin, NT, Hong Kong
Center of Computational Astrophysics
Flatiron Institute
New York City, NY, USA
hschan@phy.cuhk.edu.hk

Siu-Hei, Cheung
Department of Physics
The Chinese University of Hong Kong
Sha Tin, NT, Hong Kong
Center of Computational Astrophysics
Flatiron Institute
New York City, NY, USA
shcheungpeter@link.cuhk.edu.hk

V. Ashley Villar
Department of Astronomy & Astrophysics
Institute for Computational & Data Sciences
Institute for Gravitation and the Cosmos
The Pennsylvania State University
University Park, PA, USA
vav5084@psu.edu

Shirley Ho
Center of Computational Astrophysics
Flatiron Institute
New York City, NY, USA
shirleyho@flatironinstitute.org

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's Legacy Survey of Space and Time. 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 (ZTF CPVS). We encode their light curves using a convolutional autoencoder, and we use an isolation forest to sort each periodic variable star by an anomaly score with the latent space. Our overall most anomalous events share some similarities: they are mostly highly variable and irregular evolved stars. An exploration of multiwavelength data suggests that they are most likely Red Giant or Asymptotic Giant Branch stars concentrated in the disk of the Milky Way. 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 of 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].

NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications

Advanced techniques to search for anomalous astrophysical events are 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 40 billion objects 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, 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. The Zwicky Transient Facility contains numerous publicly available data which serve as a good training set for big data problems in astronomy which will arise with LSST. We anticipate that our pipeline can be applied once the observational data from the Vera Rubin Observatory (or other, new facilities) are made available.

2 Method

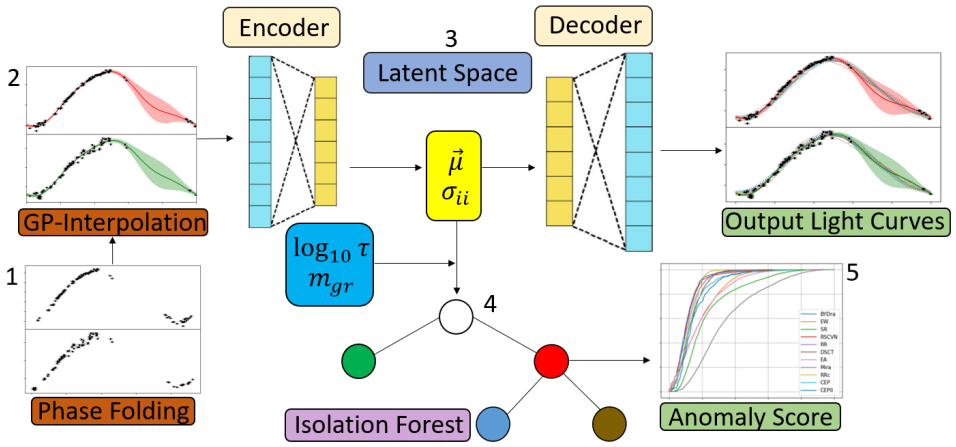


Figure 1: The anomaly detection pipeline: 1. Phase-folding the raw detection data. 2. Interpolation using the MGPR. 3. Encoding to get latent vectors $\vec{\mu}$ and matrix elements σ_{ii}. 4. Append $\log_{10}\tau$ and $m_{\rm gr}$ to $\vec{\mu}$. 5. Run the isolation forest and rank the anomalies.

Here we describe our training set and methodology. Our training set consists of the ZTF CPVS presented in Chen *et al.* [17]. The ZTF CPVS utilizes the Data Release 2 archive of The Zwicky Transient Facility [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 CPVS 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 the multivariate Gaussian process [MGPR, 19, 5, 20]) with periodic boundary conditions. The MGPR 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, $\vec{r} = (\vec{\phi}, \vec{\lambda})$ is a high dimensional position vector, while 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. After fitting the kernel function, 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]. We select $730,184$ of data and split them into a train to validation to test ratio of $7:2:1$. 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 bottleneck 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 are applied. We train with early stopping with an epoch of 895, which takes roughly 6 hours running on the NVIDIA RTX2080-Ti GPU.

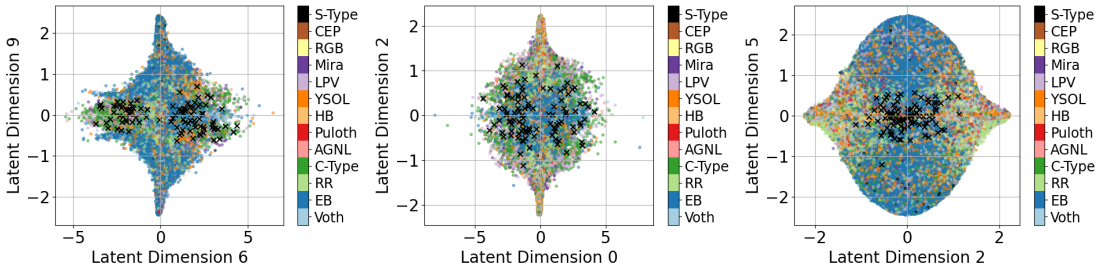
Once trained, we next use an isolation forest to rank each periodic variable star by its anomaly score. The isolation forest works by building an ensemble of binary trees, which work to isolate samples of the population. The anomalies are identified as requiring few trees to isolate the event [24]. In 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 $100,000$[1]. The use of the isolation forest is inspired by its simplicity and robustness. We noticed that there are detection scores specialized for deep-generative networks, such as the Mahalanobis confidence score [26–29], which computes distances between data points to a distribution (assumed Gaussian). It is not applicable in our case, because we have appended extra variables into our feature space in which they are not necessary Gaussian. Furthermore, we are ultimately interested in the relative *rankings* but not the absolute scores of the anomalies. A simple yet robust isolation forest should meet our expectations. We remark that reconstruction scores are also widely used as indicators for anomalies [30, 31], but our empirical findings suggest that they perform roughly similar to the isolation forest. Finally, we encourage readers to refer to Figure 1 as a review of our pipeline.

We additionally use our learned latent space to *classify*[2] the ZTF CPVS by their physical origin. We note that the ZTF CPVS provides class labels for PVSs, but the classifications are based on hand-engineered features only. Here we provide an alternative classification method that makes use of a hierarchical random forest classifier. We extracted events labels from the SIMBAD catalog [32] by cross-matching (using the python package `Astroquery` [33]) their sky-coordinates with those listed in the ZTF CPVS. The SIMBAD catalog contains class labels obtained, typically, through spectroscopic analysis, which is more reliable but often expensive. We found $31,541$ of successfully cross-matched objects. 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), other Pulsating Variables ($\rm Pul_{oth}$), RR Lyrae (RR), and Peculiar Variables (Pec). The second level is 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 Other Variables ($\rm V_{oth}$). They will be serving as the data set of our classification model. We split the data set into a training-to-test set ratio of $7:3$ by using the python package `scikit-learn` [34]. We note that our training set is highly imbalanced, with the largest set containing $10,745$ events and the smallest containing just 41 events. We balanced the training set using the python package `imbalanced-learn` [35] with default learning parameters, which performs synthetic minority resampling [36, 35]. Finally, we train the hierarchical random forest classifier provided by `imbalanced-learn`, with no hyperparameter tuning performed.

[1] We found such value to be sufficient to yield converged scores.

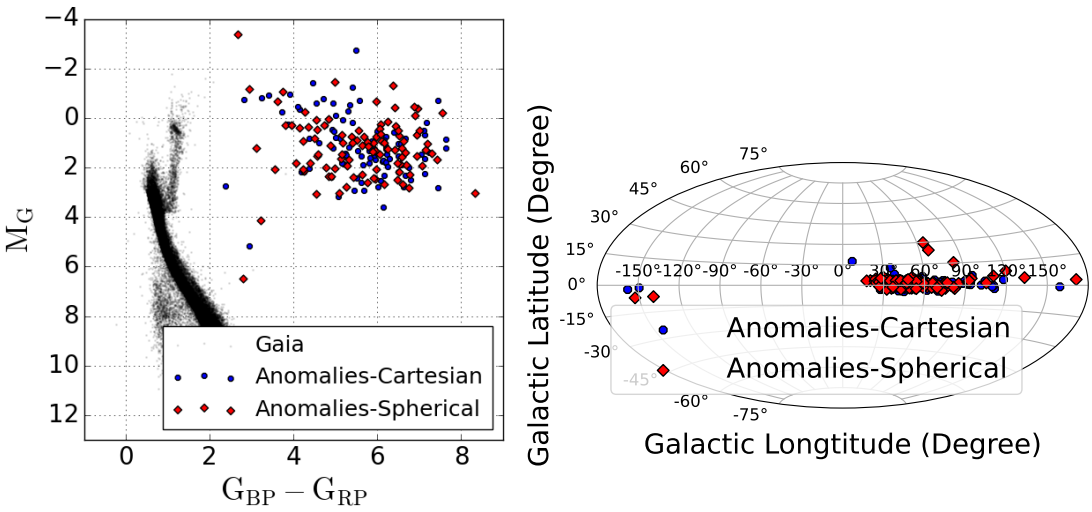
[2] In addition to the latent space, we also append hand-engineered features, including the joint period, and the amplitude and mean magnitudes of both the g- and r- band light curves.

3 Results and Discussion



(a) (b) (c)

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

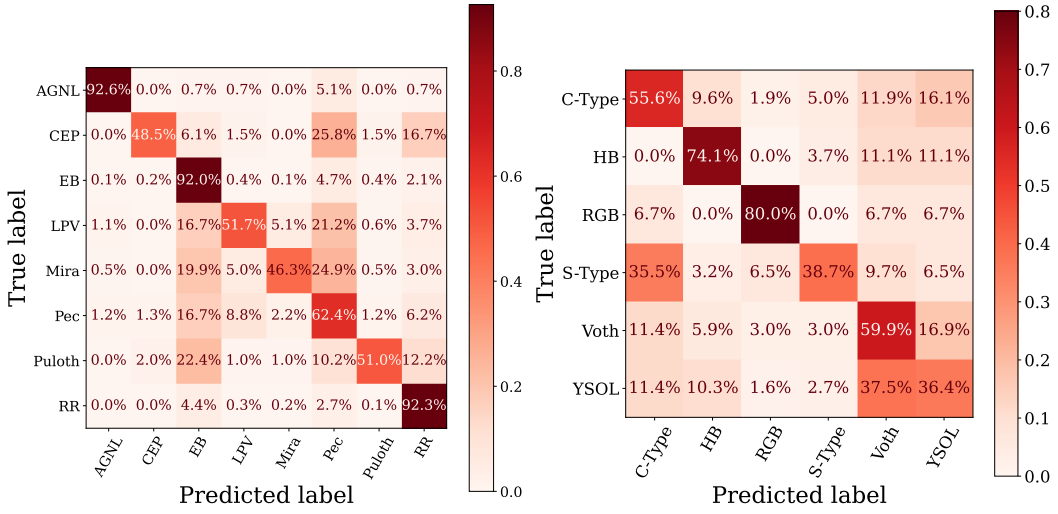


(a) (b)

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.* [37], Bailer-Jones *et al.* [38], and Gaia Collaboration *et al.* [39], through the Vizier Catalogue [40] by `Astroquery` [33]. (b) Same as (a), but for their distributions in the Milky Way galactic coordinates.

We find that our latent space exhibit an annular structure, which inspires us to transform the latent vector into an 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 [39] 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 the temperature and luminosity of the stars. The majority of the anomalies are *cool*, with $G_{\rm BP} - G_{\rm RP} > 4$ and *luminous* with $0 < M_{\rm G} < 2.5$. These properties correspond to evolved, luminous, and cold stars. Furthermore, we show the distribution of the top anomalies in the Milky Way galactic coordinates in



(a) (b)

Figure 4: (a) Confusion matrix for the first level classification labels of the test set. We show in each row the completeness of each class. (b) Same as (a), but for the second level classification labels.

Figure 3 (b). The majority of the anomalies concentrate in the Milky Way galactic disk, implying that (1) there is significant interstellar reddening due to the dust for these events and (2) the progenitor systems of these events are consistent with young and massive stars. Taken together, the observational evidence points to highly anomalous, young, cool, and massive Red Giant or Asymptotic Giant Branch stars. Spectroscopic follow-up observations and detailed light curve modeling 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 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 labeled as semi-regular *galactic* variable stars in previous studies. We hope that a closer investigation of these labels can lead to improved purity in the ZTF CPVS. We note that we do not use the same class labels as the original ZTF CPVS, making it difficult to directly use the ZTF CPVS as a baseline model. A detailed comparison between classification methods will be left to the future. Nonetheless, we find that the new classification model for the ZTF CPVS that uses our learned latent space is reasonably accurate. Last but not least, we plot the latent distribution using our new labels in Figure 2, 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) create a new classification model for PVS. 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 big data has become a rapidly growing field in recent years. We anticipate our method can contribute to the community, such as detecting other kinds of anomalous periodic/non-periodic transients or better classifying different categories of astronomical objects. Furthermore, our unsupervised learning pipeline can make use of supervised object labels to look for categorical-wise anomalies, which could potentially help better understand the diversities within classes and improve the labeling and classifications of stellar objects. Techniques similar to those presented here can be used in broader applications to identify anomalies in periodic/non-periodic, multi-variate time series.

5 Acknowledgement

We thank the Flatiron institute for providing computer cluster access. We thank Prof. Ming-Chung, Chu in the Chinese University of Hong Kong for providing the NVIDIA RTX2080-Ti GPU as the computational resources for the neural network training. We also thank Prof. Maria Drout and Anna O'Grady at the University of Toronto; Matteo Cantiello, Mathieu Renzo, and Adam Jermyn at the Center of Computational Astrophysics, Flatiron Institute for their valuable discussion on the anomalous periodic variables stars. VAV acknowledges support by the Simons Foundation through a Simons Junior Fellowship (#718240) during the early phases of this project, as well as support in part by the NSF through grant AST-2108676.

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] R. Lin, E. Khalastchi, and G. A. Kaminka, in *2010 IEEE International Conference on Robotics and Automation* (2010) pp. 3038–3044.

[27] X. Jin, E. W. M. Ma, L. L. Cheng, and M. Pecht, IEEE Transactions on Instrumentation and Measurement **61**, 2222 (2012).

[28] K. C. Ho, S. Harris, A. Zare, and M. Cook, in *Detection and Sensing of Mines, Explosive Objects, and Obscured Targets XX*, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9454, edited by S. S. Bishop and J. C. Isaacs (2015) p. 94541B.

[29] R. Kamoi and K. Kobayashi, arXiv e-prints , arXiv:2003.00402 (2020), arXiv:2003.00402 [stat.ML] .

[30] J. An and S. Cho (2015).

[31] T. Wang, X. Xu, F. Shen, and Y. Yang, IEEE/CAA Journal of Automatica Sinica **8**, 1296 (2021).

[32] 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] .

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

[34] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, Journal of Machine Learning Research **12**, 2825 (2011).

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

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

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

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

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

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