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

Ho-Sang Chan[1,2], Siu-Hei Cheung[1,2], Ashley Villar[3], Shirley Ho[2]
[1]The Chinese University of Hong Kong
[2]Center for Computational Astrophysics, Flatiron Institute
[3]The Pennsylvania State University

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

Advanced techniques to search for anomalous astrophysical events are essential in the era of upcoming observatories. Here, we provide a convolutional autoencoder based pipeline for

1. *Anomaly detection*
2. *Classifications*

of periodic variable stars (PVSs) detected with Zwicky Transient Facility (ZTF). ZTF contains numerous publicly available data which serve as a good training set for big data problems that arise in the future.

Data Pre-processing

We extracted light curves from the ZTF catalog of periodic variable stars (ZTF CPVS). The data are given in two bands of filters. We phase-folded the light curves and interpolated them using the method of **multivariate Gaussian process regression** (MGPR):

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

Here, $\vec{r} = (\vec{\phi}, \vec{\lambda})$ is a high-dimensional vector, K is the covariance function, ϕ is the phase, λ is the wavelength of the band filters, C is a constant, and δ measures white noises.

Latent Features Extractions

We generate 160 data points along the phase direction for both bands of data and stack them into an "image" of size 2×160. They were then fed into a **convolutional variational autoencoder** to generate their latent representation. The structure is:

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

The train-to-val-to-test ratio is $7:2:1$. We perform rough grid searches to optimize the hyper-parameters.

Hierarchical Random Forest

We additionally use our learned latent space to classify the ZTF CPV by their physical origin. We extracted events labels from the **SIMBAD** catalog by their sky coordinates. They were fed into a **hierarchical random forest**. We balanced the training set using the python package **imbalanced-learn** with default learning parameters, which performs synthetic minority resampling. Finally, we train the **hierarchical random forest** classifier using a training-to-test set ratio of $7:3$, with no optimization.

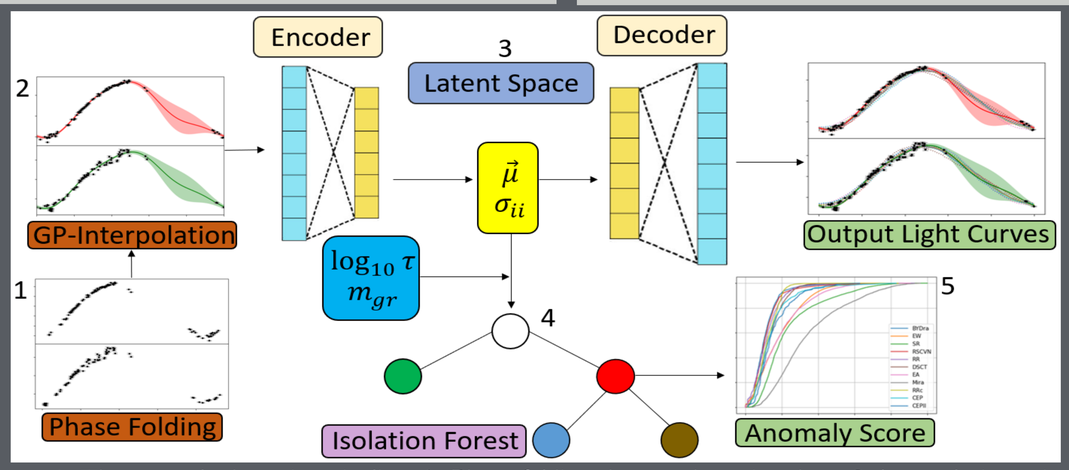

Figure 1. The anomaly detection pipeline: 1. Phase-folding the raw detection data. 2. Interpolation using the Multivariate Gaussian Process. 3. Encoding to generate latent features. 4. Append additional hand-engineered features. 5. Isolation forest and ranking the anomalies.

Results and Discussion

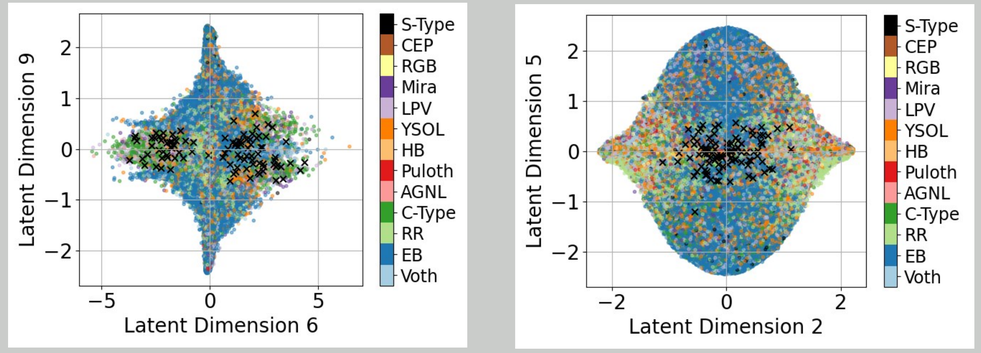

Figure 2. Examples of 2D-projection for latent distributions. Distinct colors represent PVSs of different categories. Descriptions of the labels shall be omitted. Anomalies are marked as black crosses.

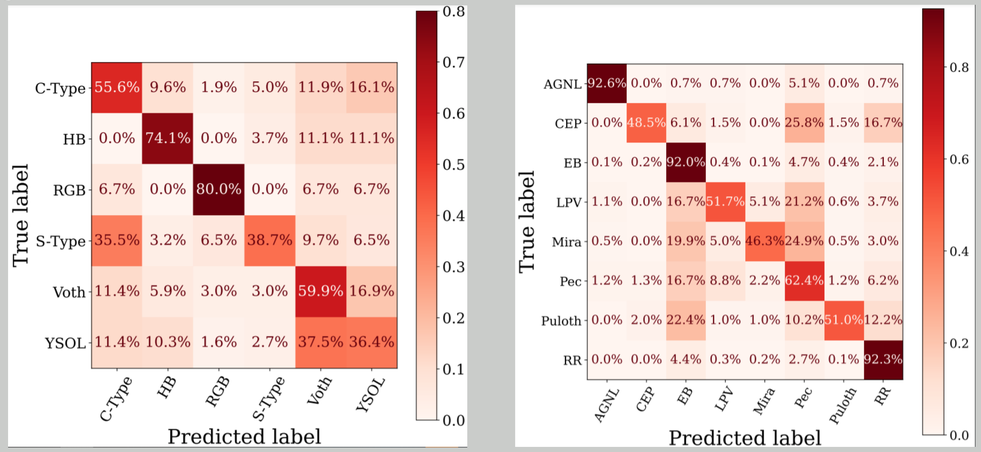

Figure 3. Confusion matrix of our new classification model. We show the completeness of each class. The description of class labels are omitted.

We ran an isolation forest with a based-estimators of 100000 in the latent space to rank each periodic variable star by its anomaly score. Our pipeline is seemingly sensitive to high-variability and irregular oscillating light curves. We found top anomalies consistent with young and massive evolved stars. We recommend spectroscopic follow-up observations to reveal their identity.

Finally, we highlight the classification results from our hierarchical classifier in Figure 3. We find that our learned latent space is sufficient to classify the set of ZTF CPVS light curves explored here with reasonable levels of accuracy.

Conclusion

We present a convolutional autoencoder-based pipeline to classify and to search for anomalous PVSs. We anticipate our method can contribute to the community, accelerating the potential for scientific discovery.