A Convolutional Autoencoder-Based Pipeline for Anomaly Detection and Classification of Periodic Variables

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

Deep-sky surveys continue to exponentially increase our discovery rates of variable cosmic phenomena. Researchers are turning towards automated methods of anomaly detection. Here, we provide a convolutional autoencoder based anomaly detection pipeline for

1. *Anomaly detection*
2. *Classifications*

of periodic variable stars (PVSs) detected with Zwicky Transient Facility (ZTF), a pathfinder survey for Large Synoptic Survey Telescope (LSST).

Methodology

We extracted light curves of PVSs from the ZTF catalog of periodic variable stars (ZTF CPVS). The data are given in two filters (g-band and r-band) of magnitudes. We phase-folded the light curves by their joint period and chose to interpolate them using the method of **multivariate Gaussian process (MGPR).** After that, we generated 160 evenly spaced data points along with the phase direction for both bands. We stacked them horizontally to form an "image" of size 2×160. They will be encoded through a **convolutional variational autoencoder** to generate latent features. We ran an isolation forest in the latent space to rank each variable by a calculated anomaly score.

In addition, we cross-matched PVSs in the ZTF catalog with those presented in the **SIMBAD** catalog to obtain more reliable and robust class labels. We found $31{,}541$ successfully cross-matched objects. We extracted their latent features together with their class labels, and they were input into a **hierarchical random forest** to train our classifier.

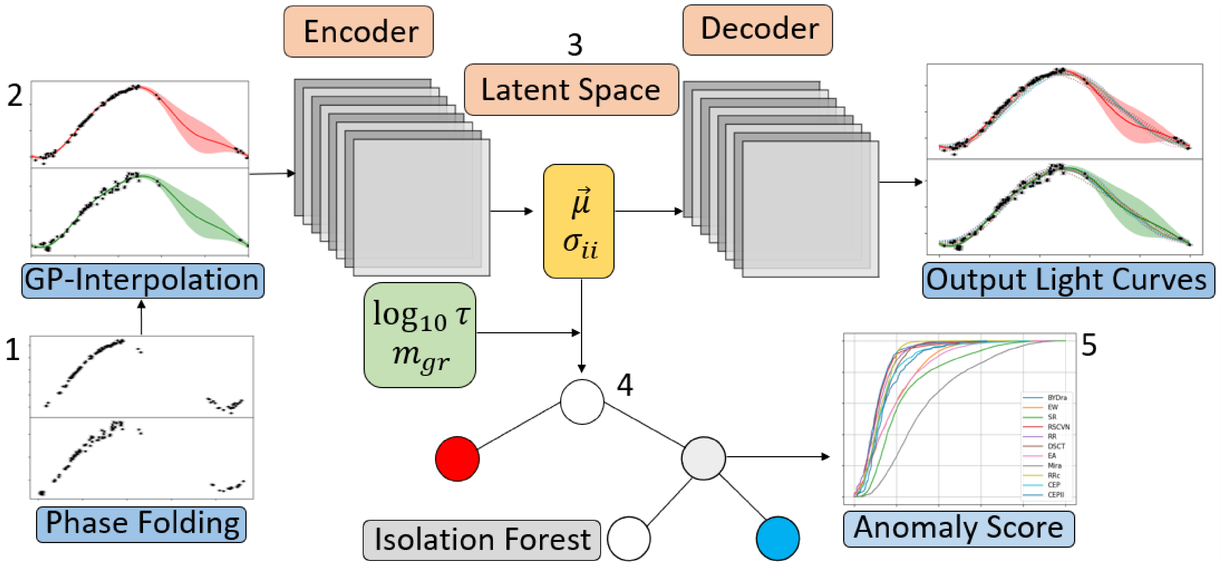

Figure 1. The anomaly detection pipeline: 1. Phase-folding the raw detection data. 2. Interpolation using the Multivariate Gaussian Process. 3. Encoding to generate latent features. 4. Append additional hand-engineered features. 5. Isolation forest and ranking the anomalies.

Top Anomalies

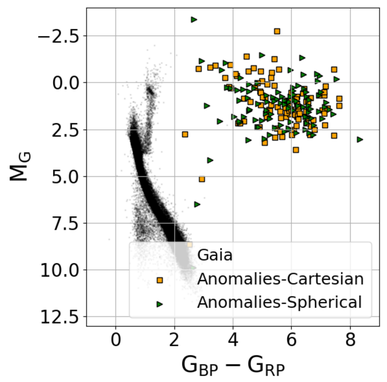

Figure 2. Distribution of the anomalies in the Gaia HR-Diagram against main-sequence stars in black dots

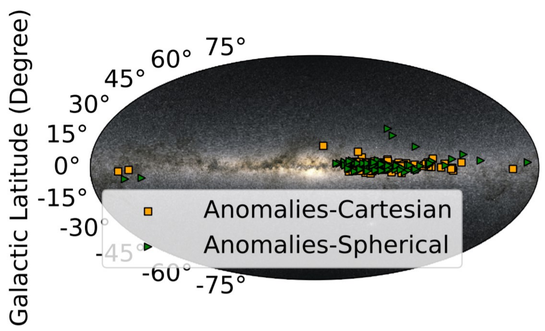

Figure 3. Distribution of the anomalies in the Milky Way galactic coordinate. Galactic longitude is omitted. Image credit: ESA/Gaia/DPAC.

We find that our learned latent space exhibits an annular structure, which inspires us to transform the latent space using **spherical coordinates**. We search for the top 100 anomalies in both latent spaces. We show their distribution in the Gaia HR-Diagram (M_G vs $G_{BP} - G_{RP}$) in Figure 2. We found that they are **bright** and **cool**, which corresponds to evolved stars. Their Galactic positions as shown in Figure 3 indicate that they are tightly located near the Galactic plane. Taken together, these stars are consistent with **young** and **massive Red Giant/AGB** stars.

The Classification Model

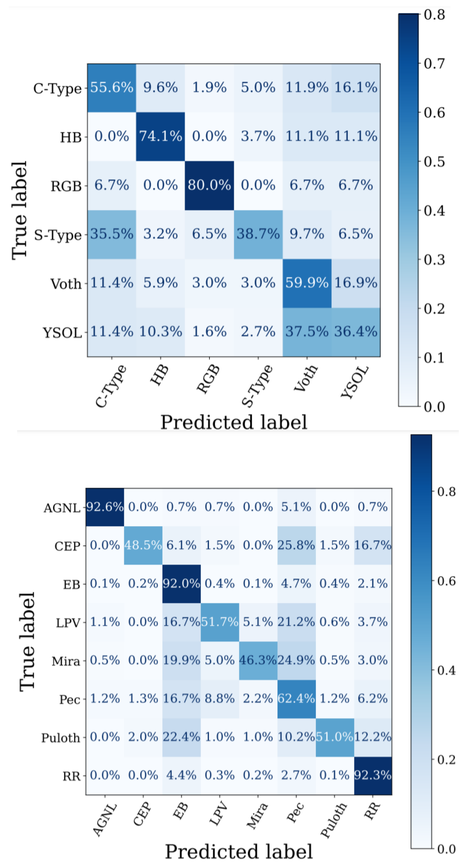

Figure 3. Confusion matrix of our new classification model. We show the completeness of each class. The description of class labels are temporarily omitted.

Finally, we highlight the classification results from our hierarchical classifier in Figure 3, where we show the completeness of each class. We find that our new classification model for the ZTF CPVS is reasonably accurate.

Conclusion

We present a convolutional autoencoder-based pipeline for anomaly detection and classification. Techniques similar to those presented here can be used in broader applications to identify anomalies in periodic, multi-variate time series.