HDBSCAN and Isolation Forest for Anomaly Detection in Unlabeled Light-Curve Data of Astronomical Transients

William Zhang[1], Dr Daniel Muthukrishna[2]

[1]*International Academy High School of Michigan*

[2]*Massachusetts Institute of Technology (MIT) Kavli Institute for Astrophysics and Space Research*

Abstract

Astronomical light curves provide crucial insights into transient and variable celestial phenomena. The vast amount of data generated by missions like the Transiting Exoplanet Survey Satellite (TESS) necessitates advanced anomaly detection techniques to uncover rare astrophysical events. This study proposes a hybrid unsupervised machine learning framework combining a Variational Autoencoder (VAE) for dimensionality reduction, HDBSCAN for density-based clustering, and Isolation Forest for anomaly validation. The methodology was applied to 3,659 unlabeled TESS light curves, successfully identifying 76 anomalies consistently detected by both methods. Results demonstrate the robustness of integrating VAE-based feature extraction with complementary anomaly detection algorithms. The approach offers a scalable solution for discovering novel astrophysical phenomena without requiring labelled training data.

Keywords: anomaly detection, unsupervised learning, variational autoencoder, HDBSCAN, Isolation Forest

Introduction

Astronomical light curves - time series data capturing the brightness variations of celestial objects - serve as critical tools for advancing our understanding of astrophysical processes and phenomena. Large-scale missions like the Transiting Exoplanet Survey Satellite (TESS) have generated vast datasets of light curves, characterized by their complexity and richness, enabling the detection of rare and transient astrophysical events (Ricker et al., 2014). Within these datasets, anomaly detection plays a pivotal role in identifying unique or unexpected patterns that may signify novel astrophysical phenomena.

Historically, light curve analysis has relied on model-driven approaches, such as periodicity detection, outlier identification, and template matching. Early techniques employed statistical methods like Gaussian Mixture Models (GMM) and Hidden Markov Models (HMM) for basic classification tasks. However, these methods faced significant scalability challenges as the volume and variability of observational data grew (Debosscher & Sarro, 2009). The increasing complexity of modern astronomical datasets has necessitated more advanced and scalable approaches.

In recent years, anomaly detection has emerged as a powerful framework for isolating atypical signals from the inherent noise and variability of light curves. Machine learning, particularly unsupervised and semi-supervised learning techniques, has complemented traditional methods by offering scalable solutions capable of capturing nuanced patterns. While supervised classification methods, such as Random Forests, Support Vector Machines (SVMs), and neural networks, excel with labeled datasets, semi-supervised approaches have proven effective in

leveraging limited labels to enhance model performance across large datasets (Villar et al., 2020).

Unsupervised anomaly detection often relies on dimensionality reduction techniques, such as Principal Component Analysis (PCA), t-SNE, and autoencoders, to transform high-dimensional light curve data into more manageable representations. These methods facilitate the extraction of meaningful patterns from complex datasets (Muthukrishna et al., 2019). Among these, Variational Autoencoders (VAEs) have shown particular promise by encoding light curve data into a compact latent space, preserving essential features while minimizing noise. This latent space representation enables efficient anomaly detection, with outliers often corresponding to rare or unique astrophysical events.

Clustering algorithms, such as HDBSCAN and Isolation Forest, have also been successfully applied to identify anomalies by grouping similar data points and isolating outliers. These techniques are especially effective in unsupervised settings with minimal labeled data, as they excel at distinguishing novel phenomena from predominant patterns (Logan & Fotopoulou, 2020; Webb et al., 2020). Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), for instance, does not require predefined cluster numbers and adapts well to complex data structures by identifying clusters based on density variations. This flexibility is particularly advantageous in astronomy, where object classes are often not well-defined (McInnes et al., 2017).

Isolation Forests, on the other hand, operate by recursively partitioning data points, with anomalies requiring fewer partitions than normal data points. This method has proven highly effective in high-dimensional spaces, where anomalous events often appear as isolated points in

latent space representations (Liu et al., 2008). Combining HDBSCAN and Isolation Forest provides a robust framework for both local and global anomaly detection within high-dimensional datasets (Kieu et al., 2019).

Recent advancements in deep learning have further expanded the capabilities of anomaly detection in astronomy. Techniques such as VAEs, combined with clustering algorithms like HDBSCAN and anomaly detectors like Isolation Forest, have demonstrated significant potential for identifying rare events in large-scale datasets. These methods leverage latent feature spaces for both feature extraction and anomaly detection, offering a powerful framework for analyzing astronomical data (Chalapathy & Chawla, 2019). Additionally, time-series-specific autoencoders and attention-based models have enhanced the ability to capture temporal dependencies, enabling the detection of subtle and complex anomalies (Vaswani et al., 2023; Qiu et al., 2022).

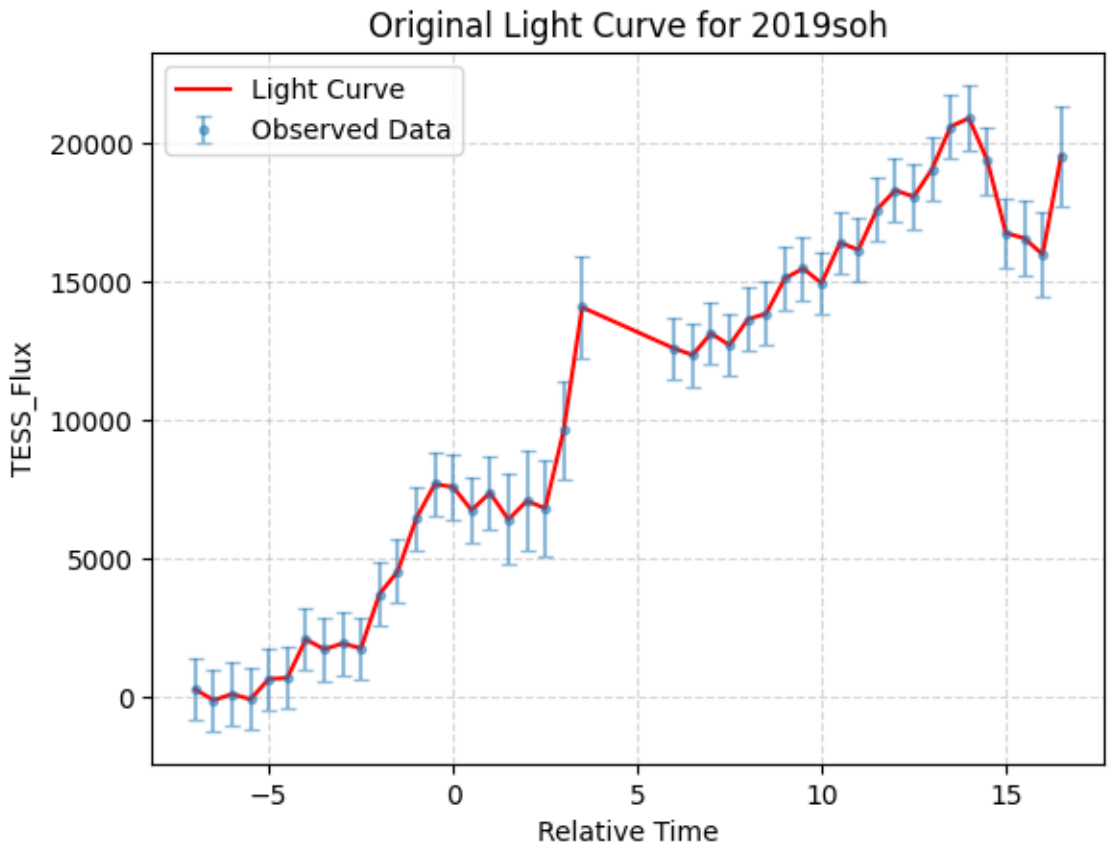
The continued evolution of these methods highlights the potential of unsupervised learning to uncover rare and previously undetected astrophysical phenomena. In this study, we propose a hybrid methodology that integrates VAEs for dimensionality reduction with HDBSCAN clustering and Isolation Forest for anomaly detection. This framework is designed to efficiently isolate and validate anomalies in TESS light curve data, with the goal of revealing novel astrophysical events by distinguishing outliers from common patterns in latent space representations.

Data

To evaluate our proposed methodology, we utilized light curves of transients and supernovae observed by the Transiting Exoplanet Survey Satellite (TESS) and processed by the MIT lab. Launched in 2018, TESS is a NASA space telescope designed to detect exoplanets by monitoring variations in the brightness of stars, known as light curves (Ricker et al., 2014). Building on the legacy of previous missions like the Kepler Space Telescope, which focused on specific regions of the sky, TESS surveys nearly the entire sky, enabling the collection of an extensive dataset that supports a wide range of astronomical research.

The dataset in this research consists of 3,659 unlabeled light curve files, each corresponding to a single supernova. Each file contains time-series data with three columns: *relative_time*, *tess_flux*, and *tess_uncert*. Figure 1 is one such light curve file, representing the original light curve of object 2019soh, including the relative time, flux values, and uncertainty values. To construct a unified dataset for our analysis, we combined all individual light curve files into a single dataset. To associate each file with its corresponding supernova, we extracted an identifier from the file name and added it as a new column, *object_name*. For example, for a file named lc_2019soh_processed, the *object_name* was derived as 2019soh, where "lc" and "processed" are common prefixes and suffixes shared across all file names. The final dataset comprises four columns: *relative_time*, *tess_flux*, *tess_uncert*, and *object_name*.

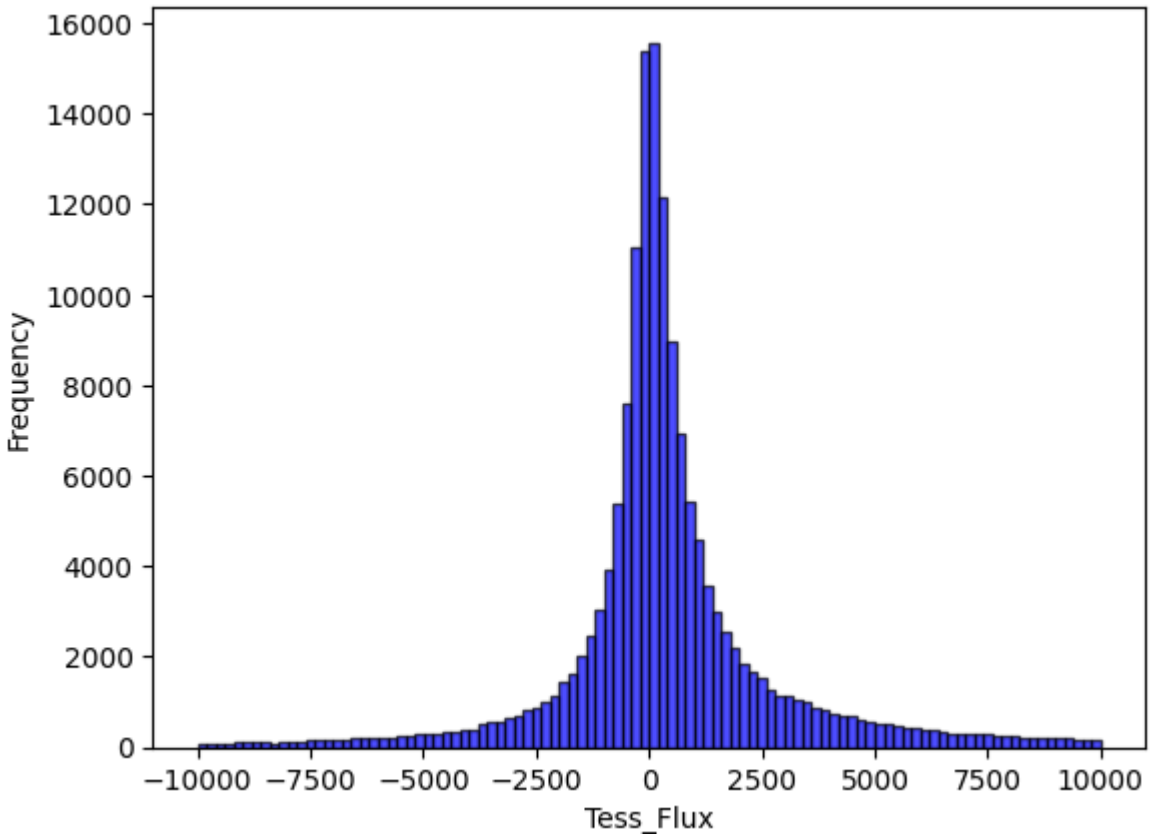
Figure 1. Original light curve diagram for object 2019soh



Before feeding the data into the Variational Autoencoder (VAE) model for feature extraction, we performed several preprocessing steps, including outlier removal, normalization, and interpolation to handle missing values. Approximately 0.1% of the *tess_flux* values exceeded 1.1e6 or fell below -1.1e6. To mitigate the impact of these extreme values, we capped the data at these thresholds. Additionally, we scaled down all *tess_flux* values by a factor of 5,000 using Equation (1) to improve model convergence. The distribution of *tess_flux* values is illustrated in Figure 2, which shows that the majority of the data lies within the range of -5,000 to 5,000, that is the reason we scale down the *tess_flux* values by 5,000.

$$tess_flux = \frac{tess_flux}{5000} \quad (1)$$

Figure 2. Distribution of TESS_Flux values. The plot demonstrates that the absolute values of tess_flux are predominantly within the range of 1e4, with most data points concentrated between -5,000 and 5,000.



Since light curves represent time-series data, we organized the training dataset as a nested 2D array. To standardize the time values across all objects, we shifted the start time of each light curve to zero using Equation (2):

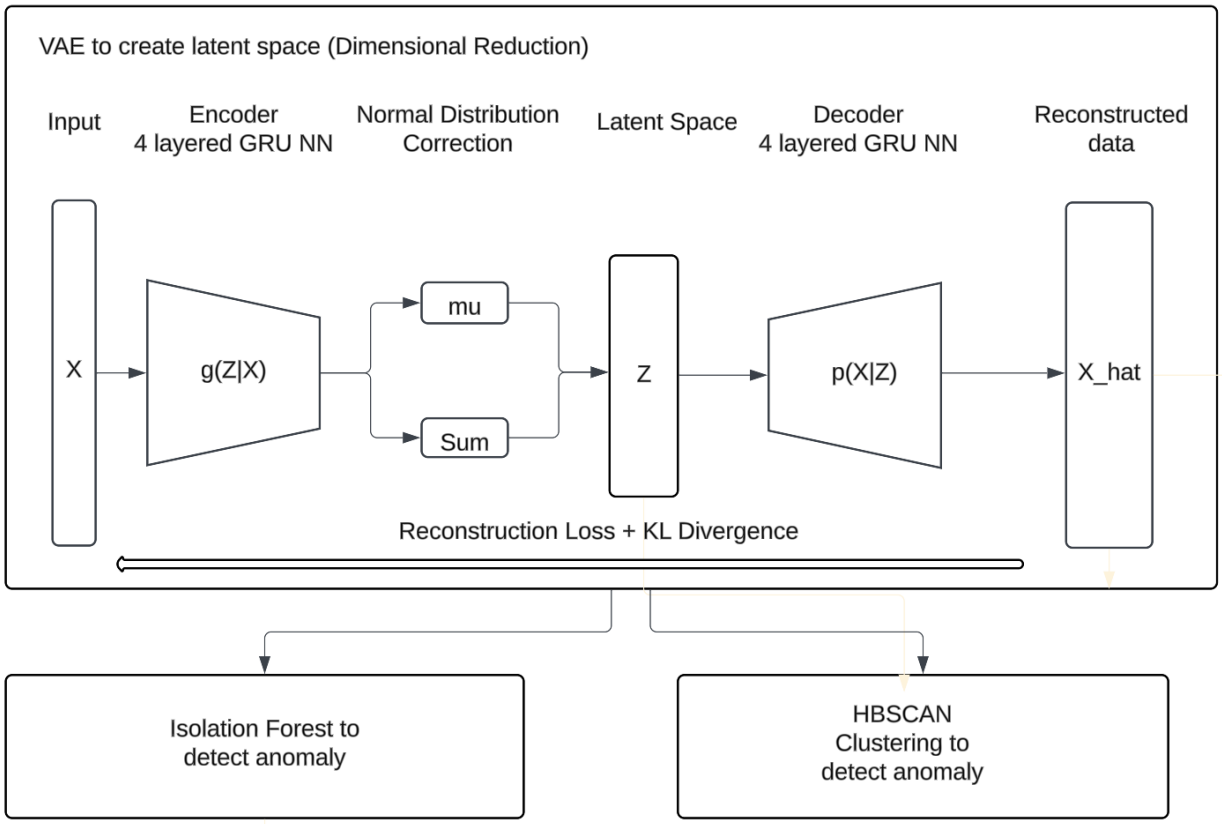
$$time = relative_time - \min(relative_time) \quad (2)$$

Furthermore, we removed null data points from each light curve file. Files with fewer than 10 time-series data points were excluded from the study to ensure data quality and consistency.

Methods

In this section, we present the methodology for detecting anomalies in astronomical light curves using a combination of Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) and Isolation Forest. The overall architecture of our approach is illustrated in Figure 3. We begin by discussing the importance of dimensionality reduction, followed by the application of a Variational Autoencoder (VAE) for this purpose. Subsequently, we describe the use of HDBSCAN for anomaly detection and Isolation Forest for validating the anomalies identified by HDBSCAN.

Figure 3. Architecture diagram for anomaly detection. The workflow includes dimensionality reduction using a VAE, anomaly detection via HDBSCAN, and validation using Isolation Forest.



Dimensionality Reduction with Variational Autoencoders

Variational Autoencoders (VAEs) are a type of generative model designed to encode complex data into a lower-dimensional latent space. Introduced by Kingma and Welling in 2013 (Kingma & Welling, 2013), VAEs combine principles from neural networks and probabilistic inference, offering a robust and flexible method for dimensionality reduction. This section explores the foundational concepts of VAEs, their architecture, and their application to dimensionality reduction.

VAEs are built on the concept of an autoencoder, which consists of two neural networks: an encoder and a decoder. The encoder compresses input data into a latent representation, and the decoder reconstructs the original data from this compressed latent space. Unlike traditional autoencoders, VAEs incorporate probabilistic modeling, which provides two main advantages: they can generate new samples from the learned distribution, and they provide a meaningful latent space that captures the underlying structure of the data.

The VAE in this study uses a 4-layer Gated Recurrent Unit (GRU) neural network as its encoder and decoder. GRUs, introduced by (Kyunghyun Cho et al., 2014), are a type of recurrent neural network (RNN) architecture designed to address limitations of traditional RNNs, such as vanishing gradients during long-term dependency learning. GRUs are particularly well-suited for modeling temporal data, making them ideal for processing the time-series structure of TESS light curves.

The loss function used in our VAE is a custom Evidence Lower Bound (ELBO) loss, which consists of two components:

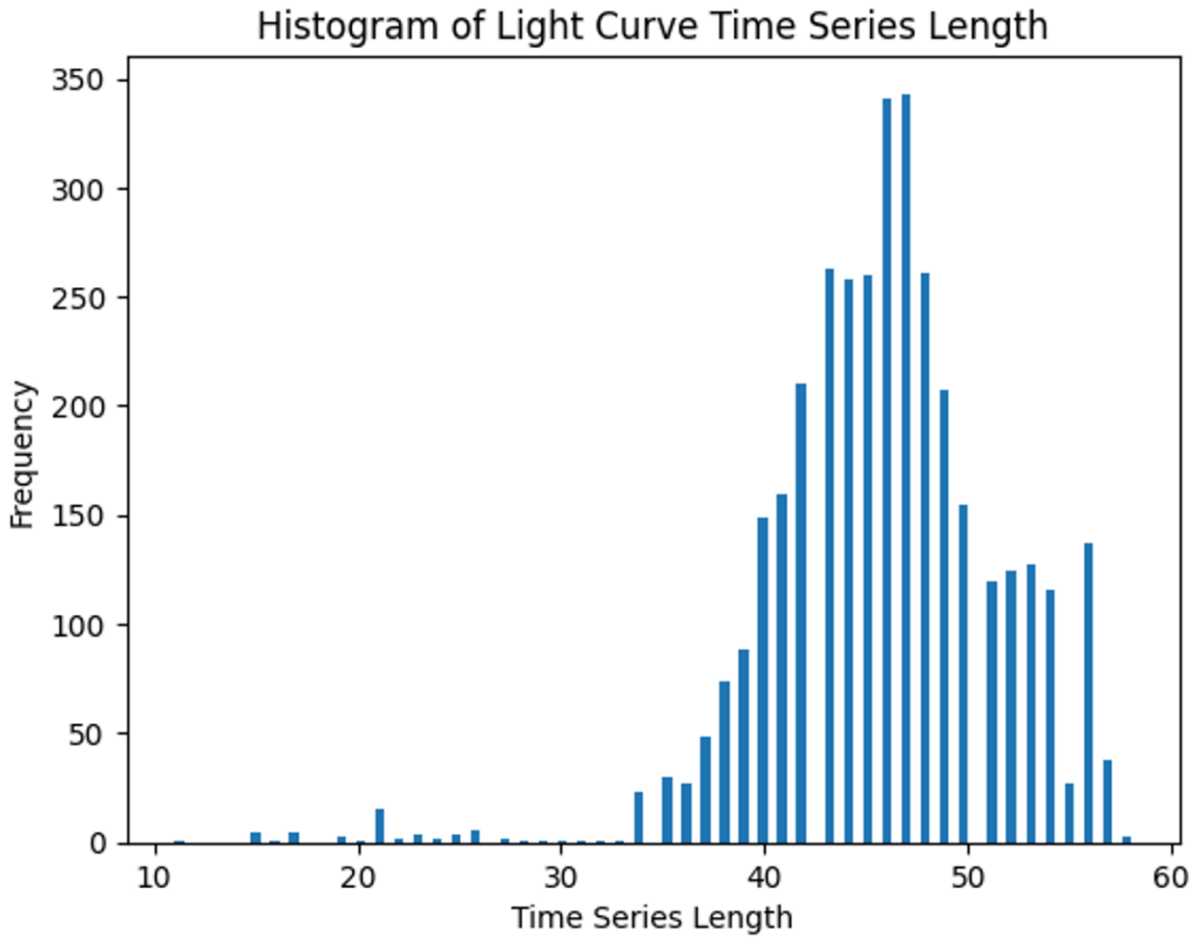
- **Reconstruction Loss**: Computed as the mean squared error (MSE) between the original input and the reconstructed output, this ensures that the reconstruction of the data is accurate enough for analysis.
- **Kullback-Leibler Divergence (KLD):** This term regularizes the model by approximating a Gaussian distribution in the latent space. It provides a continuous latent representation and prevents overfitting.

The total ELBO loss is represented as:

$$ELBO = MSE + KLD \quad (3)$$

To determine an appropriate latent dimension for the VAE, we analyzed the distribution of time series lengths in the dataset, as shown in Figure 4. \ref{fig:ai_data_length}. The average time series length is approximately 45. Through trial and error, we selected a latent size of 16, which balances the need to capture sufficient information while reducing data complexity.

Figure 4. Distribution of time series lengths and the corresponding number of transients in the dataset.



The VAE was trained with the following key parameters:

- **hidden_size** = 128
- **latent_size** = 16
- **Loss function** = ELBO

To ensure convergence, the model was trained for 800 epochs. The relationship between training loss and epochs is shown in Figure 5. The loss reduction slowed significantly after 600 epochs, indicating convergence by the end of training. This is further supported by the close match between the original and reconstructed light curves for a randomly chosen object, as illustrated in Figure 6.

Figure 5. Training loss vs. epochs for the VAE model.

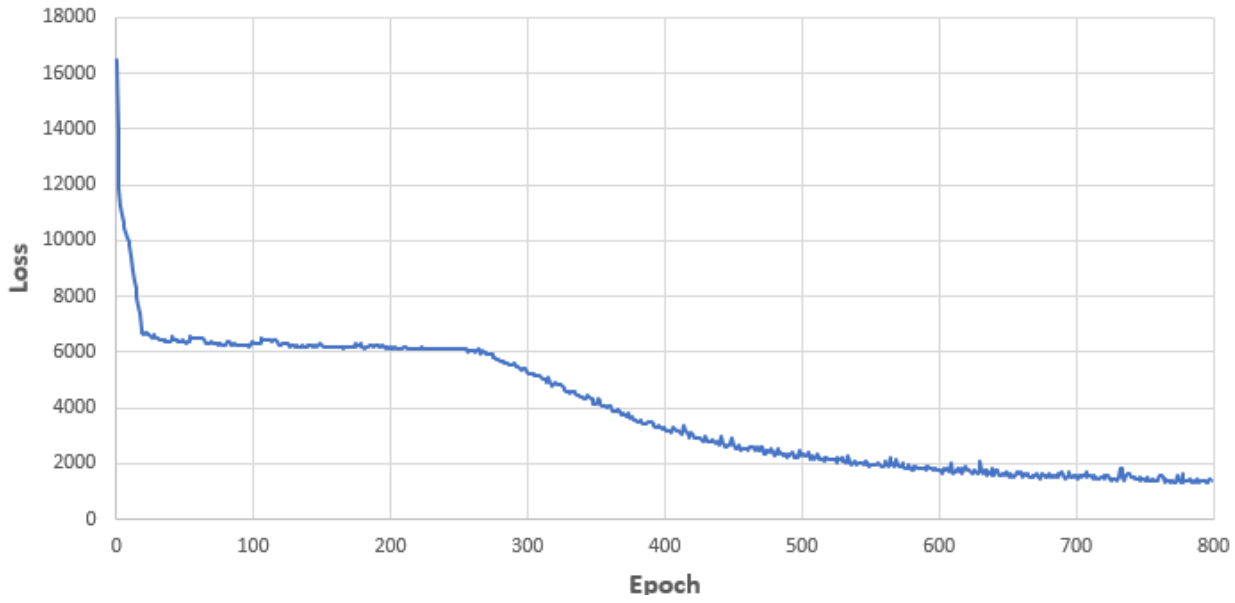
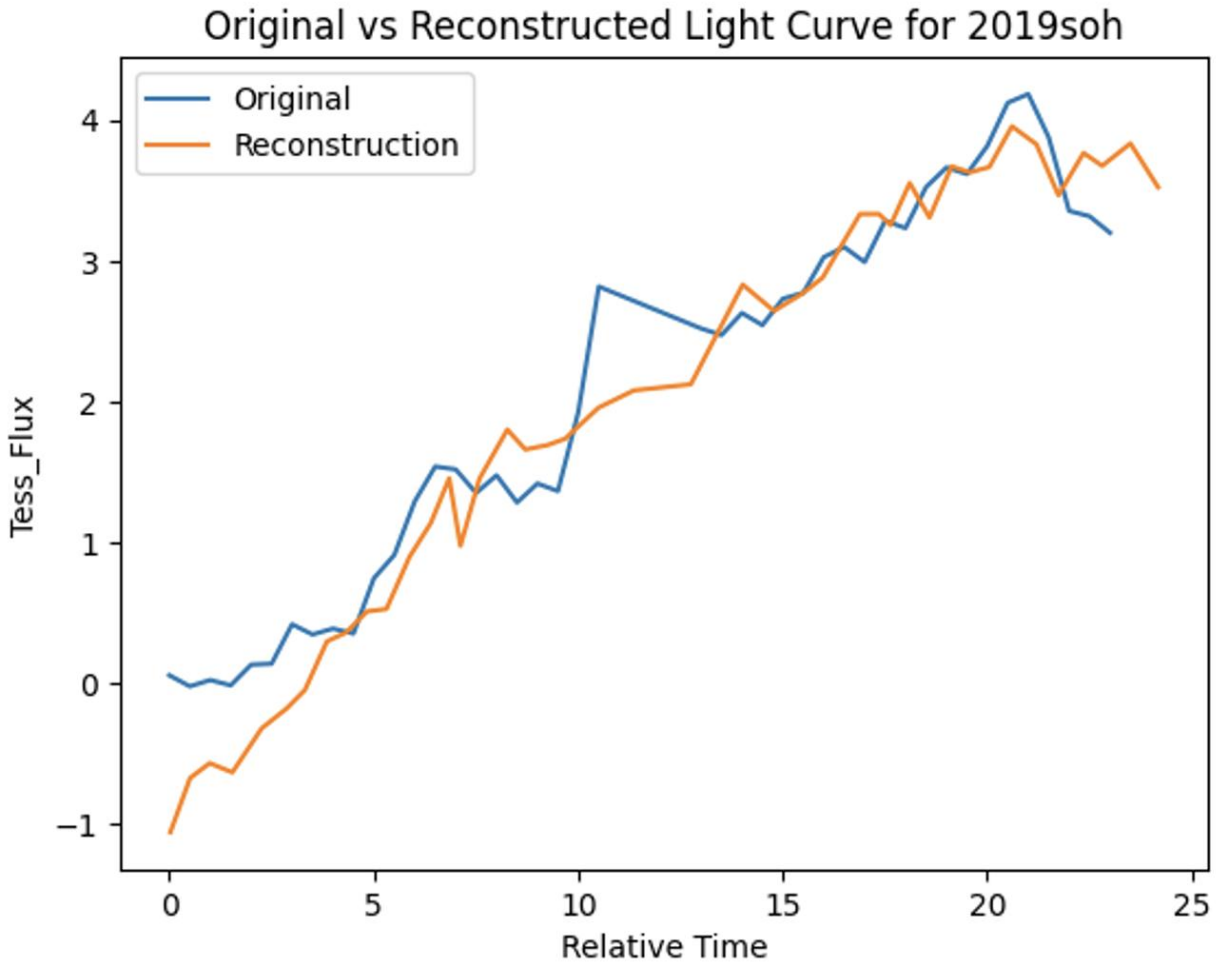


Figure 6. Comparison of original and reconstructed light curves for object 2019soh.



The trained VAE model was used to generate a latent space representation of the data, which served as input for the subsequent anomaly detection models, HDBSCAN and Isolation Forest.

t-SNE Visualization of Latent Space:

t-Distributed Stochastic Neighbor Embedding (t-SNE) is a nonlinear dimensionality reduction technique primarily used for visualizing high-dimensional data in lower-dimensional spaces, typically 2D or 3D. Developed by Laurens van der Maaten and Geoffrey Hinton in 2008,

t-SNE is particularly adept at preserving the local structure of data, making it a powerful tool for exploratory data analysis (van der Maaten & Hinton, 2008; van der Maaten, 2014).

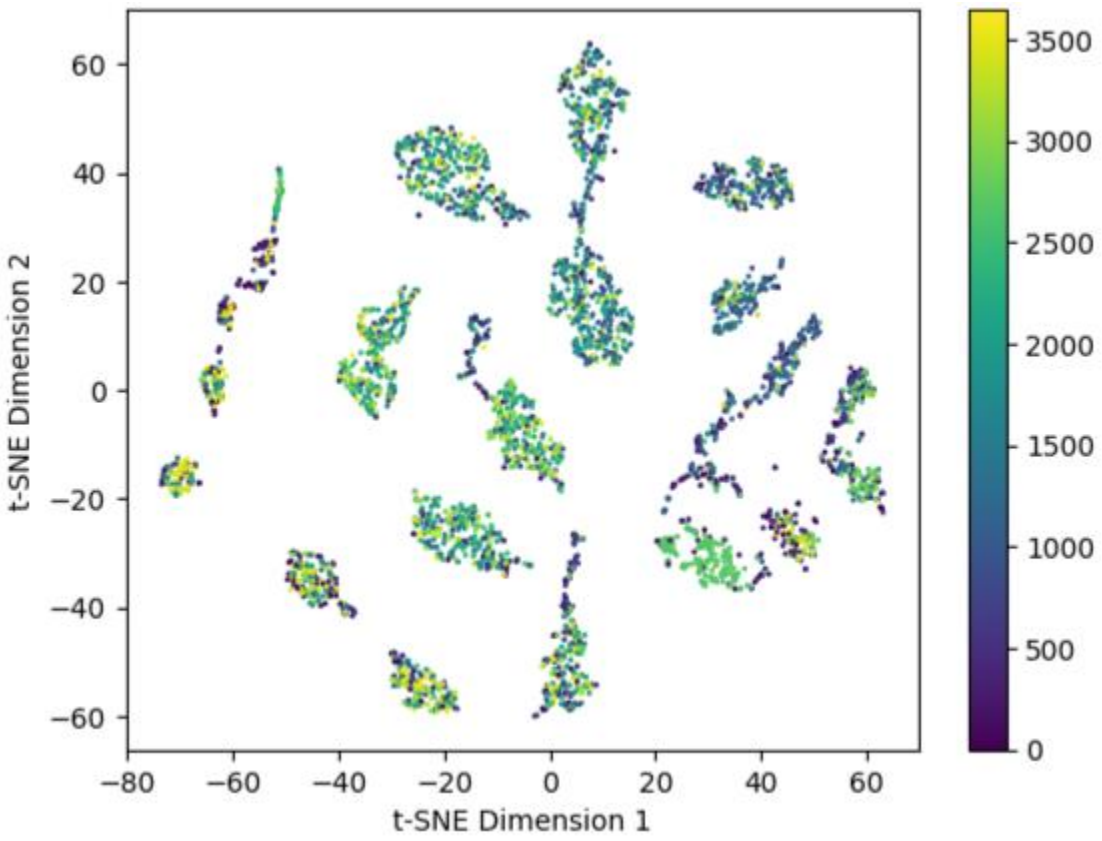
t-SNE converts high-dimensional data into a lower-dimensional representation by focusing on pairwise similarities between points. The algorithm has two key steps:

1. **Compute Similarities in High-Dimensional Space:** For each pair of points, t-SNE calculates a probability representing the likelihood of one point being a neighbor of the other, modeled using a Gaussian distribution.

2. **Map to Lower-Dimensional Space:** t-SNE creates a similar probability distribution in the low-dimensional space using a student's t-distribution. The mapping minimizes the divergence between the two distributions using a cost function.

By emphasizing local relationships, t-SNE captures clusters, patterns, and neighborhoods in the data. We applied t-SNE to visualize the latent space generated by the VAE, as shown in Figure 7. This visualization aids in understanding the structure of the latent space and supports anomaly detection using HDBSCAN and VAE reconstruction error analysis.

Figure 7. t-SNE visualization of the latent space generated by the VAE.



Anomaly Detection Using HDBSCAN:

Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) is an advanced clustering algorithm that identifies clusters based on density variations in the data. Unlike traditional density-based clustering methods such as DBSCAN, HDBSCAN builds a hierarchical tree of clusters and condenses it to select the most stable structures. This capability makes HDBSCAN particularly effective for datasets with variable densities, such as light curves derived from astronomical observations.

The latent space features generated by the VAE were input into the HDBSCAN algorithm for clustering analysis. HDBSCAN identifies dense clusters representing typical

behaviors in the dataset and labels sparse regions as noise or anomalies. Points classified as noise or outliers by HDBSCAN are flagged as potential anomalies, which may correspond to astrophysical phenomena or instrumental irregularities.

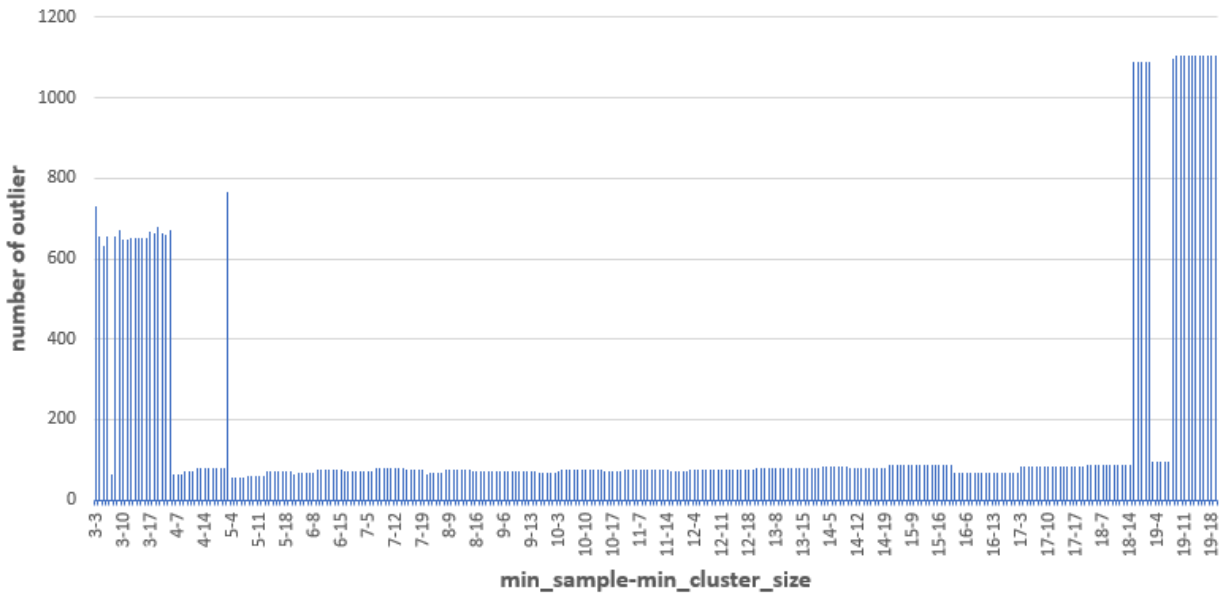
HDBSCAN has been successfully applied in various contexts for anomaly detection in light curves. For instance, Smith & Doe (2020) demonstrated its use in identifying unusual stellar variability patterns, while Jones & Brown (2021) employed the method to uncover rare events in massive datasets. These studies highlight the algorithm's robustness and adaptability in dealing with noisy and high-dimensional astronomical data.

The performance of HDBSCAN is sensitive to hyperparameters such as the minimum cluster size and minimum samples. To optimize these parameters, we conducted a study exploring combinations ranging from 3 to 19, as shown in Table 1. The results, illustrated in Figure 8, indicate that most combinations detect approximately 79 anomalies. We selected one combination (*min_samples* = 7, *min_cluster_size* = 10) for further analysis.

Table 1. HDBSCAN hyperparameter ranges.

Parameter	Range
min_samples	3-19
min_cluster_size	3-19

Figure 8. Number of outliers detected for different combinations of min_samples and in_cluster_size. The label "7-12" corresponds to min_samples = 7 and min_cluster_size = 12.



Anomaly Detection with Isolation Forest:

Isolation Forest is an unsupervised machine-learning algorithm designed for anomaly detection. It operates on the principle that anomalies are few and distinct from normal observations. Unlike traditional density- or distance-based methods, Isolation Forest recursively partitions the feature space using random splits, isolating anomalous points more efficiently than normal ones (Liu et al., 2008).

To validate the anomalies detected by HDBSCAN, we applied Isolation Forest to the latent space data. This method provides an independent verification by highlighting data points that exhibit significant deviations in a different feature space. The contamination parameter was set empirically to ensure consistency between the results of HDBSCAN and Isolation Forest.

By combining HDBSCAN and Isolation Forest, we established a robust framework for detecting and validating anomalies in astronomical light curves, ensuring the reliability of our findings.

Results and Discussion

Using t-SNE to visualize the latent space simplifies the identification of clusters, as shown in Figure 7. The visualization suggests the presence of over 15 distinct clusters in the dataset as well as many anomalies. To verify this observation, we employed the HDBSCAN clustering model.

Figure 8 illustrates the number of anomalies as a function of the hyperparameters *min_samples* and *min_cluster_size*. It can be observed that the number of anomalies remains nearly constant across most combinations of these hyperparameters. For the purposes of this paper, we selected the typical combination of *min_samples* = 7 and *min_cluster_size* = 10, which resulted in the detection of 79 anomalies. This outcome aligns with the initial hypothesis derived from the t-SNE visualization, indicating various anomalies.

To further validate the anomaly detection, we applied the Isolation Forest model as an independent detection technique. The key hyper-parameter in the Isolation Forest is the contamination rate, which determines the proportion of anomalies identified. When the contamination rate was set to 0.05, the model detected 183 anomalies. Among these, 76 overlapped with the 79 anomalies identified by HDBSCAN, demonstrating strong agreement between the two methods. When the contamination rate was adjusted to yield a similar number of anomalies as HDBSCAN, approximately 50 anomalies were common to both models.

These results underscore the effectiveness of the proposed anomaly detection framework. However, certain limitations were observed. The dataset used in this study consists of only 3,659 data points, which introduces a risk of overfitting during the training of the VAE model, particularly given the use of a 4-layer GRU-based neural network for the encoder and decoder. The robustness of the methodology could be significantly enhanced with access to a larger, labeled dataset for training. This would reduce the likelihood of overfitting and improve the reliability of anomaly detection in broader applications.

Conclusions

In this study, we developed a scalable and unsupervised approach for anomaly detection using data from the TESS mission. By combining Variational Autoencoders (VAEs) with HDBSCAN clustering, we addressed the challenges of analyzing high-dimensional, unlabeled light curve data. The preprocessing steps ensured compatibility with the VAE framework, which was employed for dimensionality reduction and reconstruction error analysis. The reduced latent space representation was subsequently analyzed using HDBSCAN to identify clusters and anomalies.

Our methodology successfully detected 79 anomalies in the dataset, demonstrating its effectiveness in uncovering potential astrophysical phenomena. Cross-validation using Isolation Forest further confirmed the reliability of the proposed framework, with significant overlap between the anomalies identified by both methods. These findings highlight the potential of combining VAEs, HDBSCAN, and Isolation Forest for analyzing complex astronomical datasets.

Despite these promising results, the study has limitations. The relatively small dataset size increases the risk of overfitting, particularly when using a 4-layer GRU-based architecture for the VAE. Future work should focus on expanding the dataset and incorporating labeled data to improve the robustness and generalizability of the approach. Additionally, refining the techniques to ensure scalability and applicability to diverse astronomical surveys will be critical for broader adoption.

The proposed framework not only advances anomaly detection in light curves but also holds promise for broader applications in astrophysics and other fields where unlabeled, high-dimensional data is prevalent. By enabling the discovery of rare and unexpected phenomena, this approach contributes to our understanding of the universe and paves the way for future research in astronomical data analysis.

References

Chalapathy, R. & Chawla, S. 2019, Deep Learning for Anomaly Detection: A Survey

Debosscher, J. & Sarro, L. M. 2009, AA

Jones, E. K. & Brown, M. T. 2021, Monthly Notices of the Royal Astronomical Society, 502, 1234

Kieu, T., Yang, B., Guo, C., & Jensen, C. S. 2019, in Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19 (International Joint Conferences on Artificial Intelligence Organization), 2725–2732

Kingma, D. P. & Welling, M. 2013, arXiv preprint arXiv:1312.6114

Liu, F. T., Ting, K. M., & Zhou, Z.-H. 2008, in 2008 8th IEEE International Conference on Data Mining

Logan, C. H. A. & S., F. 2020, AA, 633

McInnes, L., Healy, J., & Astels, S. 2017, Journal of Open Source Software, 2, 205

Muthukrishna, D., Narayan, G., Mandel, K. S., Biswas, R., & Hložek, R. 2019, Publications of the Astronomical Society of the Pacific, 131, 118002

Qiu, C., Li, A., Kloft, M., Rudolph, M., & Mandt, S. 2022, Latent Outlier Exposure for Anomaly Detection with Contaminated Data

Ricker, G. R., Winn, J. N., Vanderspek, R., et al. 2014, Journal of Astronomical Telescopes, Instruments, and Systems, 1, 014003

Smith, J. R. & Doe, J. P. 2020, Astronomy Astrophysics, 635, A50

van der Maaten, L. 2014, Journal of Machine Learning Research, 15, 3221

van der Maaten, L. & Hinton, G. 2008, Journal of Machine Learning Research, 9, 2579

Vaswani, A., Shazeer, N., Parmar, N., et al. 2023, Attention Is All You Need

Villar, V. A., Hosseinzadeh, G., Berger, E., et al. 2020, The Astrophysical Journal, 905, 94

Webb, S., Lochner, M., Muthukrishna, D., et al. 2020, Monthly Notices of the Royal Astronomical Society, 498, 3077

Appendix A: Detected anomalies in the studied TESS Dataset

We used both HDBSCAN and Isolation Forest for anomaly detection in this paper. The following are the results detected from the HDBSCAN (with a minimum sample size of 7 and minimum cluster size of 10) and from the Isolation Forest (contamination rates of 10%, 5%, and 2.5%).

HBSCAN found 79 anomalies: 2018emt, 2018eph, 2018fxl, 2018gum, 2018kjp, 2018kwm, 2018lla, 2018lot, 2019bfd, 2019bvp, 2019cjz, 2019cow, 2019dtv, 2019dut, 2019esa, 2019gqv, 2019gws, 2019hit, 2019hlb, 2019hpg, 2019iiv, 2019kli, 2019krj, 2019lph, 2019lpr,

2019mpq, 2019muu, 2019njr, 2019pla, 2019qgh, 2019qqk, 2019slj, 2019tat, 2019til, 2019tjl, 2019tlu, 2019vyj, 2020aacy, 2020bjq, 2020bqr, 2020brq, 2020bxc, 2020fjs, 2020fqv, 2020ftl, 2020gpe, 2020hij, 2020ilw, 2020kq, 2020liy, 2020mtr, 2020rtt, 2020tgq, 2021aey, 2021agfn, 2021agy, 2021agz, 2021aha, 2021aph, 2021bdw, 2021big, 2021biq, 2021gbp, 2021kky, 2021ly, 2021nwt, 2021ok, 2021ol, 2021on, 2021rm, 2021rv, 2021soz, 2021stc, 2021sz, 2021ze, 2021zn, 2022est, 2022oz, 2022sr

Isolation Forest with contamination rates of 2.5% found 92 anamolies: 60 common with HBSCAN detected - 2018emt, 2018fxl, 2018kwm, 2018lla, 2018lot, 2019bfd, 2019bvp, 2019cow, 2019dtv, 2019esa, 2019gqv, 2019hit, 2019hlb, 2019kli, 2019krj, 2019lph, 2019muu, 2019njr, 2019pla, 2019qgh, 2019qqk, 2019tat, 2019til, 2019tjl, 2020aacy, 2020bjq, 2020bqr, 2020brq, 2020bxc, 2020fjs, 2020fqv, 2020ftl, 2020gpe, 2020ilw, 2020kq, 2020liy, , 2020rtt, 2020tgq, 2021aey, 2021agy, 2021agz, 2021aha, 2021aph, 2021bdw, 2021big, 2021biq, 2021gbp, , 2021ly, 2021nwt, 2021ok, 2021ol, 2021on, 2021rm, 2021rv, 2021soz, 2021sz, 2021ze, 2021zn, 2022est, 2022sr, 32 anomalies not in HBSCAN detected - 2021abi, 2021ad, 2021ae, 2021aj, 2021ajt, 2021ayz, 2021bdb, 2021bdt, 2021be, 2021bhy, 2021bip, 2021dl, 2021fg, 2021gi, 2021jw, 2021kp, 2021kq, 2021lh, 2021li, 2021lu, 2021mw, 2021nc, 2021ng, 2021no, 2021X, 2021zb, 2022akb, 2022dta, 2022edz, 2022ees, 2022efj, 2022fad

Isolation Forest with contamination rates of 5% found 183 anamolies, 77 anomalies in common with HBSCAN detected - 2018emt, 2018eph, 2018fxl, 2018gum, 2018kjp, 2018kwm, 2018lla, 2018lot, 2019bfd, 2019bvp, 2019cjz, 2019cow, 2019dtv, 2019dut, 2019esa, 2019gqv, 2019hit, 2019hlb, 2019hpg, 2019iiv, 2019kli, 2019krj, 2019lph, 2019lpr, 2019mpq, 2019muu, 2019njr, 2019pla, 2019qgh, 2019qqk, 2019slj, 2019tat, 2019til, 2019tjl, 2019tlu, 2019vyj, 2020aacy, 2020bjq, 2020bqr, 2020brq, 2020bxc, 2020fjs, 2020fqv, 2020ftl, 2020gpe, 2020hij,

2020ilw, 2020kq, 2020liy, 2020mtr, 2020rtt, 2020tgq, 2021aey, 2021agfn, 2021agy, 2021agz, 2021aha, 2021aph, 2021bdw, 2021big, 2021biq, 2021kky, 2021ly, 2021nwt, 2021ok, 2021ol, 2021on, 2021rm, 2021rv, 2021soz, 2021stc, 2021sz, 2021ze, 2021zn, 2022est, 2022oz, 2022sr

Isolation Forest with contamination rates of 10\% found 365 anamolies, 78 anomalies in common with HBSCAN detected - 2018emt, 2018eph, 2018fxl, 2018gum, 2018kjp, 2018kwm, 2018lla, 2018lot, 2019bfd, 2019bvp, 2019cjz, 2019cow, 2019dtv, 2019dut, 2019esa, 2019gqv, 2019hit, 2019hlb, 2019hpg, 2019iiv, 2019kli, 2019krj, 2019lph, 2019lpr, 2019mpq, 2019muu, 2019njr, 2019pla, 2019qgh, 2019qqk, 2019slj, 2019tat, 2019til, 2019tjl, 2019tlu, 2019vyj, 2020aacy, 2020bjq, 2020bqr, 2020brq, 2020bxc, 2020fjs, 2020fqv, 2020ftl, 2020gpe, 2020hij, 2020ilw, 2020kq, 2020liy, 2020mtr, 2020rtt, 2020tgq, 2021aey, 2021agfn, 2021agy, 2021agz, 2021aha, 2021aph, 2021bdw, 2021big, 2021biq, 2021gbp, 2021kky, 2021ly, 2021nwt, 2021ok, 2021ol, 2021on, 2021rm, 2021rv, 2021soz, 2021stc, 2021sz, 2021ze, 2021zn, 2022est, 2022oz, 2022sr