X-Ray Source Classification Using Machine Learning: A Study with EP-WXT Pathfinder LEIA

Xiaoxiong Zuo[1,2,3] ![ORCID], Yihan Tao[1,3] ![ORCID], Yuan Liu[1], Yunfei Xu[1,3], Wenda Zhang[1], Haiwu Pan[1], Hui Sun[1], Zhen Zhang[1,2,3], Chenzhou Cui[1,3], and Weimin Yuan[1]

[1] National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100101, China; y.tao@nao.cas.cn, liuyuan@nao.cas.cn
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] National Astronomical Data Center, Beijing 100101, China
Received 2024 April 1; revised 2024 July 2; accepted 2024 July 10; published 2024 August 13

Abstract

X-ray observations play a crucial role in time-domain astronomy. The Einstein Probe (EP), a recently launched X-ray astronomical satellite, emerges as a forefront player in the field of time-domain astronomy and high-energy astrophysics. With a focus on systematic surveys in the soft X-ray band, EP aims to discover high-energy transients and monitor variable sources in the universe. To achieve these objectives, a quick and reliable classification of observed sources is essential. In this study, we developed a machine learning classifier for autonomous source classification using data from the EP-WXT Pathfinder—Lobster Eye Imager for Astronomy (LEIA) and EP-WXT simulations. The proposed Random Forest classifier, built on selected features derived from light curves, energy spectra, and location information, achieves an accuracy of approximately 95% on EP simulation data and 98% on LEIA observational data. The classifier is integrated into the LEIA data processing pipeline, serving as a tool for manual validation and rapid classification during observations. This paper presents an efficient method for the classification of X-ray sources based on single observations, along with implications of most effective features for the task. This work facilitates rapid source classification for the EP mission and also provides valuable insights into feature selection and classification techniques for enhancing the efficiency and accuracy of X-ray source classification that can be adapted to other X-ray telescope data.

Key words: methods: data analysis – X-rays: binaries – stars: variables: general – X-rays: bursts

1. Introduction

X-ray observations are important sources for the study of time-domain astronomy. Transient sources within this field, such as supernovae, gamma-ray bursts (GRBs), active galactic nuclei (AGNs), and X-ray binaries (XRBs), undergo substantial and sudden changes in radiation across the X-ray and gamma-ray spectra. Within the contemporary landscape of multi-wavelength and multi-messenger time-domain astronomy, monitoring celestial events in the X-ray range holds great scientific promise. Notably, numerous X-ray satellites, such as Swift (Burrows et al. 2005), XMM-Newton (Jansen et al. 2001), and Chandra (Weisskopf et al. 2000), are actively investigating the universe, producing a wealth of significant scientific discoveries. Time-domain phenomena are mostly characterized by their sporadic and transient nature. Rapid detection and timely monitoring of time-domain astronomical events are essential for their study and analysis. Efficient processing of large data sets is crucial in time-domain astronomy research, Furthermore, the integration of multi-wavelength data to reveal complex patterns and comprehensive understanding has resulted in a notable paradigm shift, with machine learning techniques emerging as prominent and influential tools.

Lobster Eye Micro-Pore Optics (MPO) is an innovative X-ray focusing technology known for its wide field of view and impressive imaging capabilities (Angel 1979; René 2010). The Einstein Probe (EP), utilizing MPO technology, is a dedicated astronomical satellite designed for time-domain astronomy and high-energy astrophysics (Yuan et al. 2018b). Launched in January 2024, EP is equipped to perform rapid, high-frequency, and systematic surveys of the soft X-ray sky in the time-domain (Yuan et al. 2015, 2018a). The EP mission will enable rapid detection and precise localization of transient and variable sources, as well as the acquisition of high-quality light curves and spectral data. EP consists of 12 Wide-field X-ray Telescopes (WXTs) covering the 0.5–4.0 keV range, accompanied by a Follow-up X-ray Telescope (FXT) that operates from 0.3 to 10 keV (Zhang et al. 2022b). In 2022 July, the Lobster Eye Imager for Astronomy (LEIA) was launched as the pathfinder for the WXT component of EP to verify its on-orbit performance and refine the operational parameters of the instrument. LEIA, equipped with a full-fledged WXT that

offers an extensive field of view measuring $18°.6 \times 18°.6$, successfully completed its orbital tests (Zhang et al. 2022a). It has obtained large-field X-ray measurement data for numerous celestial objects, revealing new transient sources (Sun et al. 2023; Yang et al. 2023), including the discovery of LXT 221107A (Li et al. 2022; Ling et al. 2022). The observational data gathered by LEIA establish a solid foundation and invaluable experience for the EP mission.

EP is positioned to collect a significant amount of time-domain sky survey data, primarily consisting of light curves and energy spectra. Employing artificial intelligence (AI) methodologies, including machine learning, to analyze extensive data resources provided by EP has the potential to uncover hidden insights in transient and variable sources. Researchers have developed a target detection framework using machine learning on the image data obtained by the Lobster Eye Telescope. This framework has been tested using EP-WXT simulation data, demonstrating promising accuracy and efficiency (Jia et al. 2023).

Since then, machine learning has attracted significant attention and achieved success in the automated classification of X-ray transient sources. This is evident from its application to data from XMM-Newton, Chandra, and other X-ray satellites. McGlynn et al. (2004) pioneered the application of machine learning techniques to classify X-ray sources, using oblique decision trees (Murthy et al. 1994) to categorize approximately 80,000 sources from the ROSAT survey into six distinct categories: stars, XRBs, AGNs, clusters, white dwarfs and galaxies. However, the limited positional accuracy of ROSAT data, ranging from approximately $10''–30''$, led to significant confusion in classification results. Lo et al. (2014) utilized a supervised learning approach to automatically classify 2XMMi-DR2 data obtained from the XMM-Newton mission. Tranin et al. (2022) applied naive Bayesian methods to Swift X-Ray Telescope (XRT) and XMM-Newton data. Additionally, Zhang et al. (2021) cross-matched the 4XMM-DR9, SDSS-DR12, and AllWISE databases to extract multi-wavelength features for effective classification. Yang et al. (2022) employed Random Forest to classify Chandra Source Catalog version 2 (CSCv2) data, and developed MUWCLASS, an automated multi-band processing pipeline specifically designed for X-ray sources.

The application of AI technologies, particularly machine learning, in astronomical data classification can significantly reduce labor costs and improve classification efficiency. EP data present unique challenges due to the short exposure times and limited photon counts of single observations, making source classification and identification extremely difficult. Manual identification of the sources is labor-intensive and time-consuming, and may result in the unfortunate consequence of missing the optimal observational window for conducting follow-up observations.

The EP mission primarily aims to detect transient and variable sources. Due to the limited sampling points and photons in LEIA and EP data, and differences from X-ray telescopes such as Chandra, direct calculation of power-law distribution and periodic features is impractical. Consequently, the current X-ray source classification models and features developed for Chandra, XMM-Newton, and other data cannot be directly applied to EP data. EP is expected to accumulate a substantial amount of data; therefore, it is crucial to first conduct classification research on LEIA and EP data. To support the requirements of the EP team for single-observation classification and the discovery of new celestial objects, there is an urgent need for a machine learning classification algorithm capable of rapidly and accurately identifying transient and variable sources in real-time.

In this paper, we propose a machine learning model that classifies target sources based on statistical characteristics of light curves, energy distributions, and other relevant features utilizing simulated EP data and LEIA observational data. The classification model has been implemented as a pipeline and deployed on the LEIA data processing server, enabling fast and real-time source classification during observations. The model can also be applied to EP in the future.

This paper is structured as follows. Section 2 introduces the EP and LEIA data, presents the characteristics of simulated data, and describes the methods used for data pre-processing and data set construction. Section 3 describes the feature extraction and selection for the classifier. Section 4 provides the details of the processing methods and model optimization techniques used in this study. The performance of the developed model is presented in Section 5. Section 6 discusses contentious issues in classification models, along with their application in the pipeline. Finally, Section 7 provides a summary of the study, highlighting the key findings and potential implications for the field of astronomy.

2. Data

The data set utilized in this research comprises LEIA observational data and simulated EP data. The data set is accessible within the China-VO PaperData Repository.[4] Both the LEIA observational data and the simulated EP data encompass a variety of file types, including the catalog, spectrum, and light curve. The event file contains information about photon arrival times, photon energies, and the positional coordinates at which the photons intersect the detector plane. The catalog file serves as a high level data product of EP-WXT, containing information extracted from detected sources in a CMOS detector. This information includes counts, pixel positions, and celestial coordinates, and is stored in the catalog file as an index-ordered list of rows within a binary table extension. The light curve file is another high-level data

---
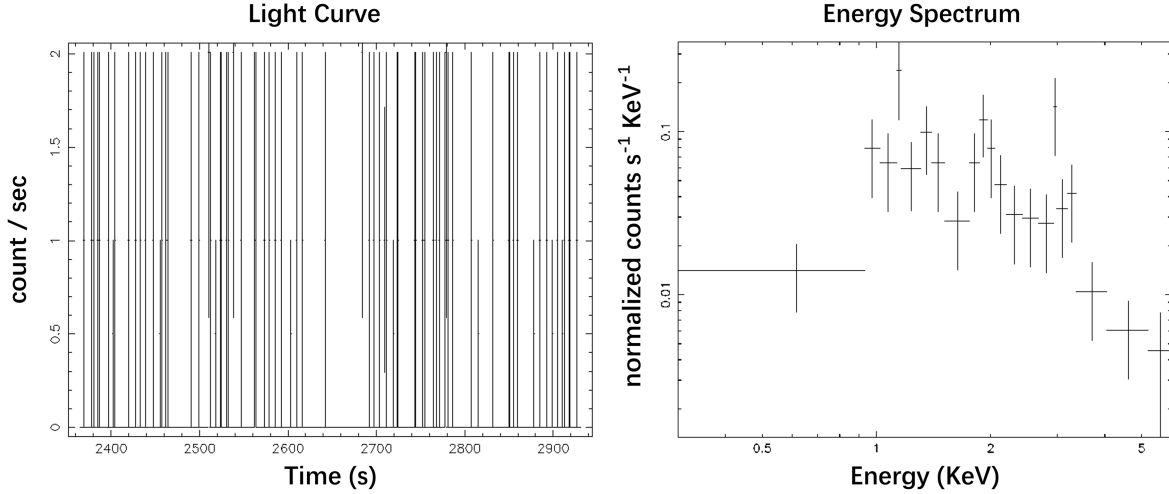
Figure 1. A quick view of an example of LEIA data, where the left panel displays the light curve and the right panel displays the energy spectrum.

product of EP-WXT generated from the event file of the EP pipeline. It provides the light curve for WXT, consisting of photon counting rates with a time resolution of 1 s. The spectrum files provide a record of the distribution of photon counts within the energy range of 0.5–4.0 keV. Figure 1 displays the light curve and energy spectrum of a low mass X-ray binary as observed by LEIA.

2.1. LEIA Data

LEIA, as the pathfinder of EP-WXT, was responsible for carrying the complete WXT test module into orbit. LEIA has an $18.°6 \times 18.°6$ field of view, an angular resolution of $3.'8$–$7.'5$, four CMOS sensors, a bandpass of 0.5–4.0 keV, an effective area of 2–3 cm^2 per 1 keV, a pixel size of 15 μm, and a total of $4k \times 4k$ pixels (Zhang et al. 2022a).

By 2023 August, LEIA had carried out 9063 observations, with 8172 of them undergoing manual verification. The data are obtained from the EP Time Domain Astronomical Information Center (TDIC),[5] which provides functionalities for data querying and downloading. Artifacts were systematically removed from our data set, resulting in categories such as AGNs, XRBs, stars, galaxy clusters, pulsars, supernova remnants (SNRs), and others. The data obtained by LEIA are derived from the certification conducted by the TA team after the operational activities of LEIA. The currently known source categories were cross-validated with source tables from other satellites, allowing for the assignment of classification labels to the observed sources. Figure 2 illustrates the histogram of LEIA single observation durations, revealing that the data are relatively short, spanning a few hundred seconds to over a

thousand seconds. Table 1 presents the quantity and proportion of each class of LEIA data.
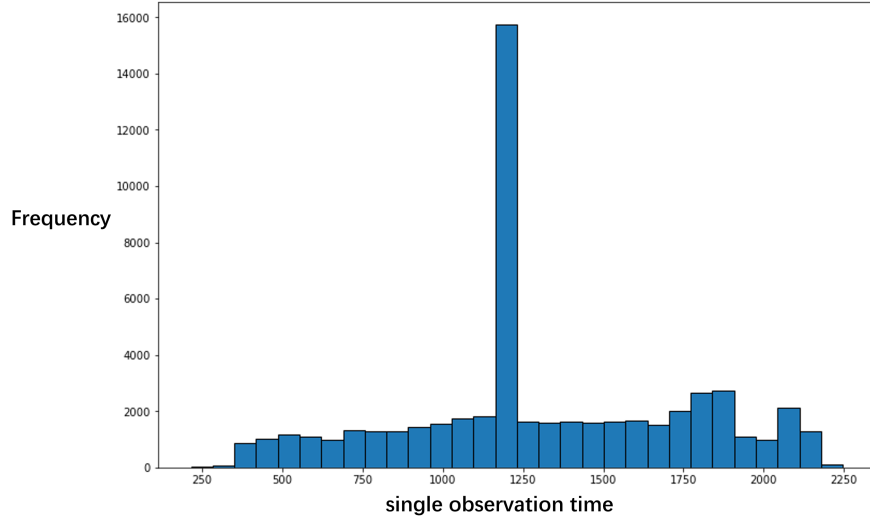
2.2. EP Simulation Data

The EP simulation data were generated using a data simulator developed by the EP Science Application Team (Pan 2024, in preparation). The simulator employs the Monte Carlo method to generate data and introduces noise into the data set. All target sources were selected from the ROSAT Skylight target directory. The simulation data are generated based on the pointing direction of EP, following the design specifications of the EP-WXT instrument. These simulated data closely replicate actual observational data and include event files, catalog files, spectrum files, light curve files, and other data types. The exposure times of simulated data range from 1100 to 1300 s, with a detector plane size of 4096×4096 pixels. The energy distribution of X-ray photons ranges from 0.5 to 4.0 keV. Figure 3 shows the histogram of EP simulation single observation durations, which primarily exceed one thousand seconds.

The simulated data are categorized into 11 types: Active Galactic Nucleus (AGN), High-Luminosity Gamma-Ray Burst (HLGRB), Galactic Star, Galactic Compact Binary Black Hole, Galactic Compact Binary Neutron Star, Galactic Compact Binary Pulsar, Galactic Compact Binary White Dwarf, Galactic Compact Binary Neutron Star and Black Hole, Short-duration Gamma-Ray Burst (SGRB), Supernova Shock Breakout (SN_SBO), and Tidal Disruption Event (TDE). To meet the requirements of the EP science team, and through simplification and consolidation, these categories were reclassified into seven types: AGN, Star, XRB, SGRB, HLGRB, SN_SBO, and TDE.
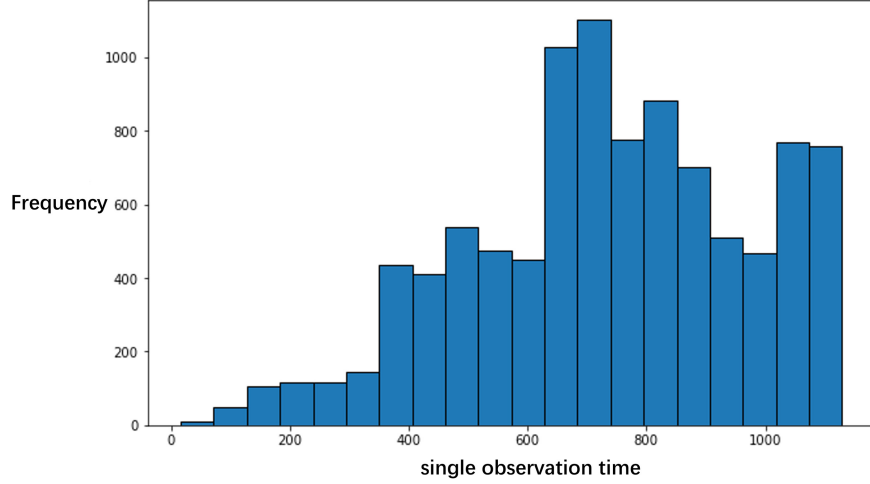
Figure 2. Histogram of LEIA single observation durations.

Figure 3. Histogram of EP simulation single observation durations.

The labels for the simulated data are derived from those assigned during the data generation process using the ROSAT catalog (Voges et al. 1999). In 1996, the X-ray source table from the ROSAT satellite sky survey was published, documenting over 18,000 X-ray sources with a positional accuracy of approximately 10″.

As a single observation in the simulated data can capture multiple sources simultaneously or no source at all, we perform source location cross-matching between the catalog file and the simulated data directory. We use a matching radius of 0°.05 and consider the matched data as a data set for classification in our study. Due to the short exposure time during observations and instrumental limitations, there are a limited number of data points in the light curve. Therefore, light curve data with fewer

than 350 data points were initially excluded. The classes of the simulated data are listed in Table 2.

2.3. Training and Test Data Set

Since the simulated data are generated based on the ROSAT star catalog, the available data categories are relatively limited. Consequently, categories such as cosmic rays and clusters of galaxies observed in LEIA are not included. However, it is important to note that LEIA has a relatively short operational period and has not yet observed rare categories such as GRBs and TDEs. These categories include rarely observed transient sources and unknown sources that are not currently present in the LEIA data but are the focus of EP's future detection efforts.

Table 1
Class Distribution and Proportion in LEIA Data

Class	Number	Proportion
X-Ray Binary (XRB)	4834	69.56%
Supernova Remnant (SNR)	949	13.66%
Cosmic ray	351	5.05%
Active Galactic Nucleus (AGN)	311	4.48%
Star	227	3.27%
Pulsar	159	2.29%
Cluster of Galaxies	118	1.70%

Table 2
Class Distribution and Proportion in EP Simulation Data

Class	Number		Percentage
	Count	Total Count	
Galactic Compact Binary Neutron Star	10674	36803	60.74%
Galactic Compact Binary White Dwarf	9242		
Galactic Compact Binary Pulsar	8435		
Galactic Compact Binary Black Hole	5452		
Galactic Compact Binary Neutron Star and Black Hole	3000		
Active Galactic Nucleus (AGN)		10662	17.60%
Galactic Star (Star)		5496	9.07%
Short-duration Gamma-Ray Burst (SGRB)		1575	2.69%
Supernova Shock Breakout (SN_SBO)		5559	9.18%
Tidal Disruption Event (TDE)		376	0.62%
High-Luminosity Gamma-Ray Burst (HLGRB)		117	0.19%

We have developed three approaches for constructing training sets for our model: using only EP simulated data, using only actual LEIA observational data, and combining EP simulated data with LEIA data. Section 5 provides a comprehensive description of the comparison among these three methods for constructing the data sets.

The final model was developed using a data set that combines EP simulated data and LEIA observational data, encompassing all categories. The training data set consists of EP simulated data and 80% of the LEIA observational data, comprising a total of 32,406 data points. The distribution of classes in this merged data set is shown in Table 3.

The data were divided in a class-balanced manner, with 80% of the combined data used for model training and the remaining 20% reserved as a mixed data test set to evaluate the model. Subsequently, the trained model is applied to the LEIA data. The remaining 20% of LEIA observational data are designated as the LEIA data test set to assess the model's performance on the LEIA data. While the model was evaluated using both test sets, particular emphasis was placed on assessing its effectiveness on the LEIA data test set.

As indicated in Table 3, our training data set exhibits a significant label imbalance. The distribution of celestial body types across the celestial sphere is uneven, leading to a scarcity of certain rare transient sources. For instance, there is an overabundance of XRBs, while the number of rare sources such as GRBs and TDEs is inadequate. The number of XRB sources is approximately 300 times greater than that of HLGRBs. These imbalances can significantly impact the performance of machine learning algorithms. To address this issue, we utilized the Synthetic Minority Oversampling Technique (SMOTE; Chawla et al. 2002) to resample the data set. This technique augmented the underrepresented categories and mitigated the problem of class imbalance. SMOTE employs the K-nearest neighbors (KNN) method to generate synthetic samples for the minority class. This approach is commonly used to address class imbalance and is recognized for its robustness.

We conducted experiments using various resampling scenarios and observed that increasing the volume of data led to improved outcomes. We attempted undersampling the categories with sufficient data, such as XRB and AGN, while resampling the remaining classes. The impact of different sample sizes, ranging from 800 to 32,000, is illustrated in Figure 4. We found that once the data quantity exceeded 16,000, there were diminishing returns in terms of accuracy and Macro-F1 scores. Consequently, our final approach involved applying SMOTE to resample all classes except for XRB. For XRB, we performed random undersampling to obtain 16,000 samples. This resulted in a total of 16,000 samples for each class.

Due to certain limitations in feature calculation, the value of some feature may be null or infinite. In such cases, we assign a uniform value of -100 to these features. Filling in missing values in this manner does not impact the model's performance.

3. Feature Extraction

The short timescale of a single EP observation poses a challenge in capturing the periodic behavior of the targets in the time domain. Additionally, instrumental limitations constrain the applicability of commonly used features, such as time variability, periodicity, power law, and flare-related features, to EP data. The extraction of features that uncover the underlying physical significance of the data is a critical task. The process of feature extraction necessitates meticulous consideration of their underlying physical meanings. The design of the feature extraction method refers to the studies conducted by Lo et al. (2014) and Richards et al. (2011). Table 4 summarizes the key characteristics of the different astronomical transients and variables considered in this study, including their timescales, light curve characteristics, and energy spectrum characteristics,
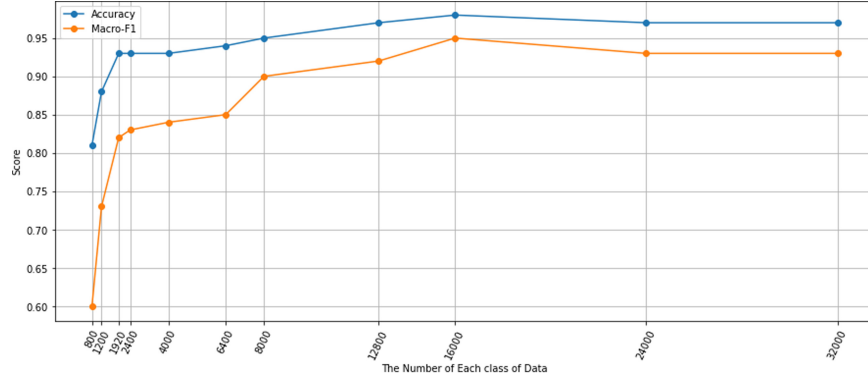
Figure 4. Accuracy and Macro-F1 under different sampling conditions of the data.

Table 3
Quantity of Data on Each Class in Different Sets

Class	Training Set	Mixed Data Test Set	LEIA Data Test Set
X-Ray Binary (XRB)	16,000	8141	985
Active Galactic Nucleus (AGN)	8735	2182	56
Star	4534	1140	49
Short-duration Gamma-Ray Burst (SGRB)	1269	306	0
Supernova Shock Breakout (SN_SBO)	455	104	0
Tidal Disruption Event (TDE)	295	81	0
Supernova Remnant (SNR)	604	154	191
High-Luminosity Gamma-Ray Burst (HLGRB)	92	25	0
Cosmic Ray	237	51	63
Cluster of Galaxies	74	21	23
Pulsar	111	25	23
Total	32,406	12,230	1390

Table 4
Characteristics of Sources and Phenomena in X-Ray Band

Class	Timescale	Light Curve Characteristics	Energy Spectrum Characteristics
AGN	Minutes to years	Aperiodic variability	Low flux density
XRB	Milliseconds to years	Strong aperiodic variability	High flux density
Cluster of Galaxies	…	No variability	Low flux density
Cosmic Ray	Milliseconds	Photons concentrated in a single readout frame	Very low flux density
Pulsar	Milliseconds to years	Periodic and aperiodic variability	High flux density
SNR	…	No variability	High flux density
Star	Kiloseconds to years	Weak variability; occasionally exhibiting a stellar flare	Low flux density
TDE	Minutes to years	Weak variability	Low flux density, Soft spectrum
SN_SBO	Minutes to hours	Transient flare	Low flux density, Soft spectrum
GRB	Seconds to minutes	Transient short-term flare	High flux density, Hard spectrum

while the energy spectrum for each class can be found in Appendix A and the light curves in Appendix B. Table 5 presents a list of the 23 features derived from the data, along with their detailed descriptions. The final selection includes the top nine features. This section delves into the design and extraction of features.

3.1. Spectral Features

The energy spectrum range detected by EP spans from 0.5 to 4.0 keV. The hardness ratio is a widely employed feature in X-ray astronomy for characterizing the spectral morphology of X-ray sources. This method involves comparing the photon counts detected in two or more distinct energy bands, typically

Table 5
List of Time Series Features Used for Classification

Feature	Description	References
galactic longitude	Galactic longitude of source	
galactic latitude	Galactic latitude of source	
a_hard	The count rates in the 0.2–0.5 keV	
b_hard	The count rates in the 0.5–1.0 keV	
c_hard	The count rates in the 1.0–2.0 keV	
de_hard	The count rates above 2.0 keV	
kurt	Kurtosis of the distribution of count rates; calculated using scipy.stats.kurtosis	
skew	Skewness of the distribution of count rates; calculated using scipy.stats.skew	Richards et al. (2011); Lo et al. (2014)
modulation index	Variance / mean	Improvement from (Lo et al. 2014)
var	Variance of the counts	
beyond1Std	Percentage of observations that lie beyond one standard deviation from the mean	Richards et al. (2011); Lo et al. (2014)
energy_ratio	The ratio between peak energy and background energy	
mean	Mean of the counts	
median	Median of the counts	
percentile_diff	Count rate at the 98th percentile minus the count rate at the 2nd percentile	Richards et al. (2011); Lo et al. (2014)
maximum slope	Maximum slope of adjacent observation points	Richards et al. (2011); Lo et al. (2014)
median_abs_deviation	Median of the absolute value of the deviation from the median	Richards et al. (2011); Lo et al. (2014)
percentage_within_threshold	Percentage of measurements within 20% of the median	Richards et al. (2011); Lo et al. (2014)
t50	50% energy width on the sides of the peak position of the energy spectrum	
t20	20% energy width on the sides of the peak position of the energy spectrum	
t10	10% energy width on the sides of the peak position of the energy spectrum	
t50_t20	t50 / t20	
t50_t10	t50 / t10	

Note. The top nine features are selected features.

categorizing them into high-energy (hard X-rays) and low-energy (soft X-rays) bands. Although the hardness ratio is a fundamental and effective technique, it may not fully capture the complexities of an X-ray spectrum.

Hardness ratios are determined by analyzing the counts in selected energy bands, which are chosen based on the specific instrument and the scientific questions under investigation. In this study, we have delineated the following energy bands: 0–0.5 keV, 0.5–1.0 keV, 1.0–2.0 keV, and above 2.0 keV. Experimental comparisons indicate that utilizing energy band counting yields superior results.

Energy band counting involves partitioning the X-ray data into distinct energy ranges and enumerating the photon counts detected in each band. This approach provides more granular information, allowing for a more comprehensive understanding of the characteristics of X-ray sources. Consequently, energy band counting enhances the robustness of the analytical algorithms employed. The energy spectrum for each class can be seen in Appendix A.

3.2. Light Curve Feature

The power-law characteristics cannot be investigated due to the brevity of the light curve. However, specific statistical features can be extracted from the available data. Below are some of the features that are essential for model training. The light curve for each class can be seen in Appendix B.

3.2.1. Kurt

Kurtosis is a statistical feature used to describe the distribution of a light curve. It quantifies the sharpness or flatness of a probability distribution curve relative to its mean. More specifically, kurtosis characterizes the steepness of the data distribution curve relative to the standard normal distribution. For our calculations, we utilize the *scipy.stats.kurtosis* function from the SciPy package (Virtanen et al. 2020).

Kurt is defined by the following equation

$$\text{kurt} = \frac{m_4}{m_2^2}. \tag{1}$$

$m2$ represents the second-order central moment (variance) of the data set, and $m4$ represents the fourth-order central moment of the data set.

3.2.2. Skew

Skewness is a statistical measure used to quantify the asymmetry of a probability distribution. In light curves, skewness indicates the degree of asymmetry in the temporal changes of luminosity. A positive skewness value suggests a longer right tail in the distribution, while a negative skewness value implies a longer left tail. In astronomy, various celestial objects exhibit diverse patterns of luminosity changes. Quantifying the skewness of a light curve enables us to
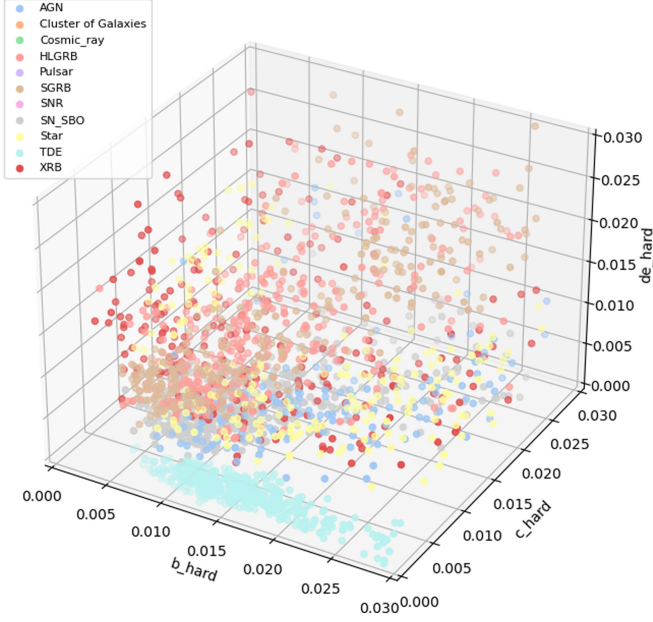
Figure 5. The distribution of each class of data in the three features b_hard, c_hard and de_hard.

understand the probability distribution characteristics of observed luminosity changes, thereby enhancing our comprehension of the underlying physical processes. For calculations, we employ the *scipy.stats.skew* function from the SciPy package (Virtanen et al. 2020).

Skewness is defined as follows

$$\text{Skew} = \frac{1}{n}\sum_{i=1}^{n} \frac{(x_i - \bar{x})^3}{s^3}. \tag{2}$$

n is the number of data points and s is the standard deviation.

3.2.3. Modulation Index

The relative volatility index can be obtained by dividing the variance of the light curve by its mean. This characteristic is referred to as the "modulation index." The relative volatility allows for the comparison of volatility among different light curves. A large relative volatility indicates significant changes in the photometric data, while a small relative volatility suggests relatively stable changes. A higher relative volatility value indicates a more active and unstable light variation phenomenon.

We randomly selected 500 data points from each class to examine their distribution in the feature space using different features, as depicted in Figures 5 and 6. It is evident that these features effectively distinguish between the classes. Figure 5 illustrates the distribution of data from different classes in three hardness ratio spaces: 0.5–1.0 keV, 1.0–2.0 keV, and above 2.0 keV. Figure 6 presents the distribution of data from

different classes based on skewness and the 0.5–1.0 keV hardness ratio.

3.3. The Distribution of the Data on the Sky Map

The types of X-ray sources can be partially distinguished by considering Galactic longitude and Galactic latitude. Previous studies have utilized Galactic latitude as a classification feature (McGlynn et al. 2004; Lo et al. 2014; Tranin et al. 2022). The distribution of Galactic longitude and Galactic latitude is also influenced by the telescope's survey design. In our study, we consider the use of Galactic longitude and Galactic latitude as effective features for classifying sources in the EP data.

Machine learning algorithms that incorporate Galactic longitude and Galactic latitude as features of position information can achieve higher accuracy rates due to the distinct position distributions of various celestial objects. By combining Galactic longitude and Galactic latitude as location information features with other attributes of celestial objects, more complex feature vectors can be constructed. Currently, the manual determination of the EP observation source also takes into account the location information for assessment. EP observations yield a relatively large number of high-energy celestial objects, including XRBs, which tend to cluster near the Galactic center. Figure 7 illustrates the spatial distribution of the data in the observed sky area.

4. Classification Methods and Procedures

All sources are labeled. We extract above described features from the data of observation sources with labels for supervised learning. This section will introduce the process of data processing, feature extraction, and machine learning in detail.

4.1. Algorithm

In this work, our primary algorithm of choice is Random Forest, an ensemble learning technique that harnesses the collective power of multiple decision trees (Breiman 2001). The core principle underlying Random Forest revolves around its combination of bootstrap aggregating and random feature selection. By employing these strategies, the algorithm aims to introduce increased randomness and diversity into the model, thereby enhancing its overall performance.

The strength of Random Forest lies in its ability to generate an ensemble prediction by aggregating the outputs of all the individual trees through either majority voting for classification tasks or averaging for regression tasks. The collective decision-making process of Random Forest leads to robust predictions that exhibit high accuracy and stability.

Random Forests are particularly well-suited for handling data sets with a large number of input features and samples. The algorithm's inherent randomness and the aggregation of multiple trees allow it to effectively capture complex
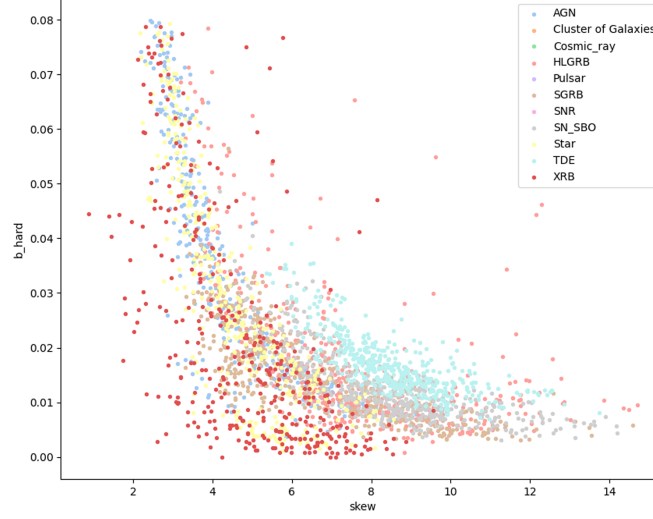
Figure 6. The distribution of data on the two features of skew and b_hard.
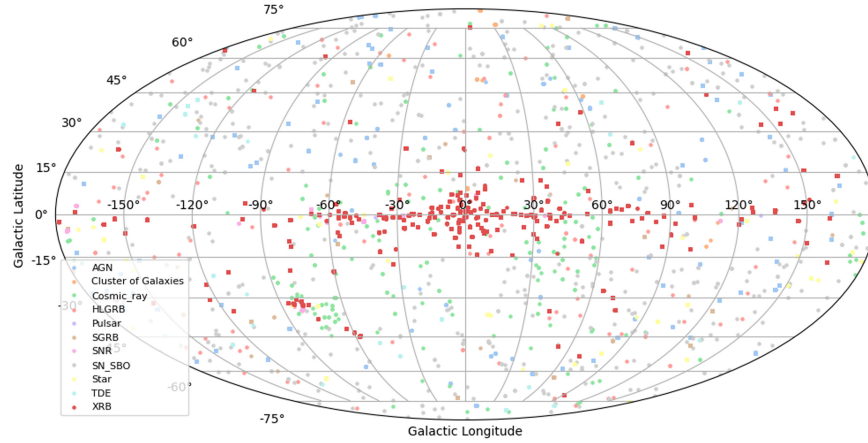
Figure 7. The distribution of the data on the sky map.

relationships and patterns within the data. Furthermore, Random Forests offer the advantage of assessing feature importance. By analyzing the contribution of each feature to the model's performance, we gain valuable insights into the key factors driving the observed patterns and outcomes.

We employed the Random Forest algorithm to learn and harness the light curve variability features and spectral features of the data, achieving remarkable results. By leveraging the power of ensemble learning, Random Forest effectively captured the intricate patterns and relationships within the data. The fusion of the data's temporal dynamics and spectral attributes within the Random Forest framework proved highly effective, with the algorithm's ability to combine the predictive strengths of multiple decision trees through voting yielding impressive outcomes.

4.2. Feature Selection

Feature importance quantifies the impact of individual features on the performance of a machine learning model. This analysis aids in identifying the most influential features, thereby enhancing model efficiency, interpretability, and our understanding of the factors driving the predictions.

In order to identify the most essential features, we initially incorporated 23 features for training the classifier. Table 5 provides a comprehensive list of these features along with their descriptions. We evaluated the contributions of features across various samples, analyzing both their individual and cumulative significance.

Due to the characteristics of the data, not all features provide equal informativeness when applied to EP data. Features with lower importance are deemed to have limited significance. To
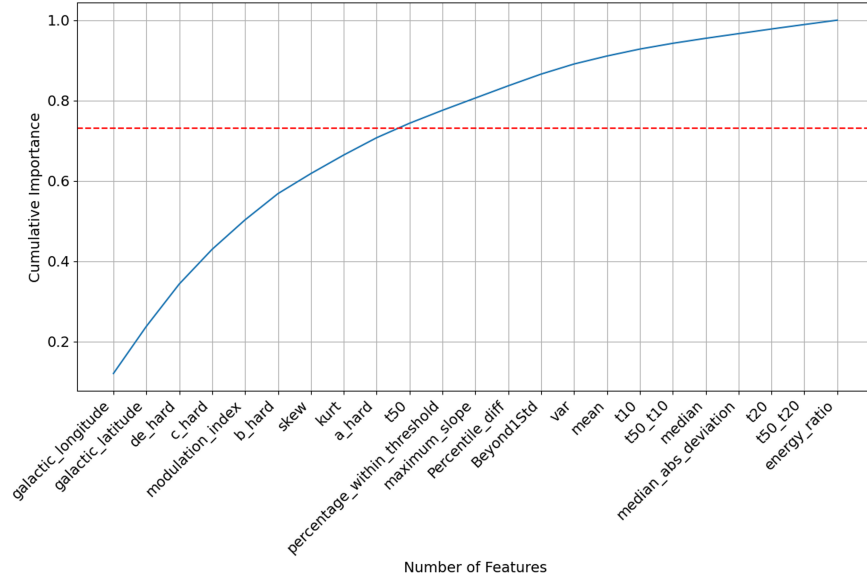
Figure 8. The cumulative importance of features. The red line is a threshold of 0.8.

Table 6
Comparison of the Effect of Feature Selection

	LEIA Data Test Set		Mixed Data Test Set	
	Accuracy	Macro-F1	Accuracy	Macro-F1
23 features	96.8%	92.3%	88.0%	83.2%
9 selected features	97.8%	94.4%	95.0%	85.4%

Table 7
Cross Validation Results

	Average	Variance
Accuracy	98.5%	6.199e-07
Macro-F1	87.0%	7.824e-05

select the most informative features, we applied a threshold based on the cumulative importance score. Figure 8 displays the cumulative importance ranking of all features, and we selected features with a cumulative importance score below 0.73.

When utilizing all 23 features, the classification performance was good, but after performing feature selection and reducing the feature set to 9, the classifier's performance improved significantly in both the mixed data test set and LEIA data test set. The classification accuracy and Macro-F1 score increased notably in both test sets. Table 6 presents the results of the feature selection comparison, highlighting the improved performance achieved by utilizing the selected subset of features.

4.3. Cross-validation

Cross-validation is a statistical technique used to assess the performance and generalization ability of machine learning models. It involves dividing the data set into multiple subsets, iteratively training the model on a portion of the data, and validating it on the remaining subsets. This process is repeated multiple times, with different subsets used for training and validation in each iteration.

By employing cross-validation, we can obtain more robust and reliable performance evaluation results, as the model is tested on multiple subsets of the data rather than relying on a single train-test split. This approach helps to mitigate overfitting and provides a more accurate estimate of the model's performance on unseen data. The results of the cross-validation are presented in Table 7.

4.4. Hyperparameter Selection

We use a cross-validated grid search to select the best hyperparameters. Grid search evaluates the performance of each parameter combination by searching for the best parameter combination in the parameter space and using cross validation. In the cross validation process, the data set is divided into 5 folds, with 1 fold used as the validation set and the other 4 folds used as the training set each time. This can comprehensively evaluate the performance of the model and reduce the impact caused by the randomness of data set partitioning.

The Random Forest algorithm consists of three main adjustable hyperparameters: the total number of trees
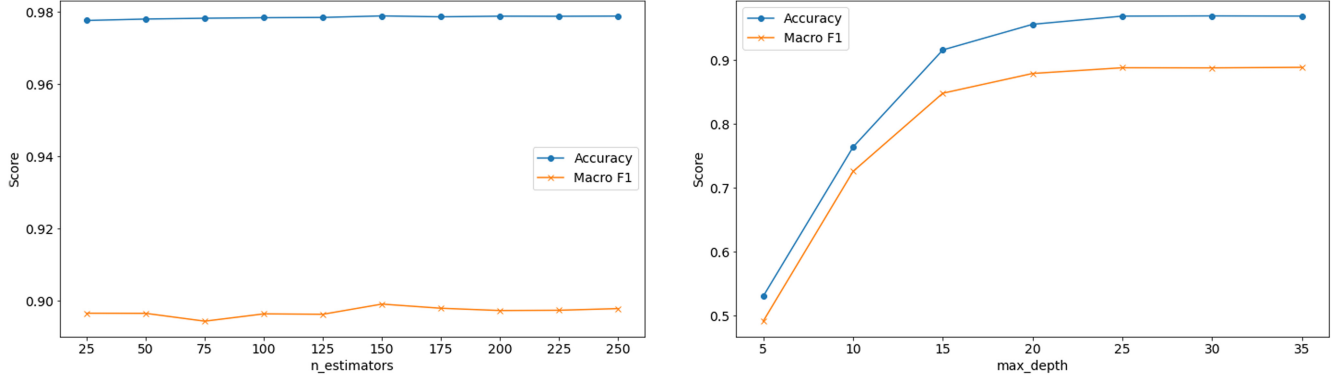
Figure 9. Figure of hyperparameter selection. The left figure displays accuracy and Macro-F1 scores for different n_estimators values, while the right figure illustrates accuracy and Macro-F1 scores for different max_depth values.

(n_estimators), the maximum depth of each decision tree (max_depth), and the maximum number of features used by each tree node (max_features). We choose the default $'auto'$ value for max_features. Using cross-validated grid search, we evaluated the hyperparameters n_estimators and max_depth while keeping the other hyperparameters fixed in their default settings. The results of the hyperparameter selection are illustrated in Figure 9. Our findings indicate that the model performs best when hyperparameters are set to n_estimators = 150 and $max\_depth = 25$.

5. Results

5.1. Evaluation Indicators

In this study, we employ five evaluation metrics to assess the effectiveness of the classification models. These metrics serve as robust indicators of model performance, including accuracy, balanced accuracy, Macro-F1 score, Matthews Correlation Coefficient (MCC), and run time.

Accuracy is calculated by dividing the number of correctly classified samples by the total number of samples.

Balanced Accuracy is calculated by averaging the accuracies of each class, resulting in a balanced accuracy indicator that effectively addresses the bias caused by data imbalance. The formula for calculating Balanced Accuracy is

$$\text{BalancedAccuracy} = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{P_i}. \qquad (3)$$

Among them, N is the number of classes, TP_i is the true number of samples in the ith class, and P_i is the total number of samples in the ith class.

The Macro-F1 score is an evaluation metric that considers both accuracy and recall. It is calculated by averaging the precision and recall values across all categories, resulting in the Macro-F1 value. This metric treats each class equally, making it robust to data imbalance.

The MCC is an evaluation indicator that provides a comprehensive measure of the relationship between true positive, true negative, false positive, and false negative predictions. It is particularly suitable for data sets with imbalanced categories. The MCC value ranges from −1 to 1, where 1 indicates perfect prediction, 0 represents random prediction, and −1 indicates completely inconsistent prediction.

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}. \qquad (4)$$

Run time is a practical metric that measures the time required for model training. It reflects the efficiency and speed of the model, making it particularly valuable when dealing with large data sets.

5.2. Algorithm Comparison

In this research paper, we attempted several popular machine learning algorithms.

XGBoost (Chen & Guestrin 2016) is a gradient boosting framework that integrates regularization and parallel processing. Compared to Random Forests, XGBoost's strategy is more focused and sequential. For our comparative study, we implemented XGBoost using the Python library *xgboost*, aligning hyperparameters with those of the Random Forest model to ensure a fair comparison. The results showed that XGBoost's performance was competitive, nearly matching the robustness of Random Forest predictions.

The KNN algorithm (Cover & Hart 1967) classifies data through the majority vote of its k nearest neighbors in the feature space. For our analysis, we used the *KNeighborsClassifier* from the sklearn library (Pedregosa et al. 2011), setting n_neighbors = 5 as the default. The algorithm uses the Minkowski distance metric and a leaf node size of 30 to

Table 8
Performance Comparison of Different Algorithms

	Random Forest	XGBOOST	KNN	GaussianNB	SVM
Accuracy	97.8%	96.5%	91.4%	28.1%	39.2%
Balanced accuracy	95.5%	95.2%	91.5%	51.9%	60.0%
Macro-F1	94.3%	81.5%	64.2%	30.5%	30.1%
MCC	95.5%	93.2%	84.6%	18.1%	36.9%
Training Time	84.902	127.022	0.283	0.048	1087.391

balance efficiency and accuracy. The uniform weighting scheme ensures equal contribution from all neighbors.

The Naive Bayes classifier is a probabilistic model that applies Bayes' theorem while assuming feature independence. We implemented the *GaussianNB* from Sklearn. Despite the model's simplicity, its assumptions of feature independence and Gaussian distribution can be restrictive, potentially affecting its performance in complex data sets where these conditions are not met. However, its effectiveness in the probabilistic classification of X-ray sources, as studied by Tranin et al. (2022), highlights its utility in specific applications.

Support Vector Machines (SVMs) (Cortes & Vapnik 1995) are a class of powerful supervised learning models known for their ability to find an optimal hyperplane that separates different classes with the maximum margin. In our predictive model, we employed the sklearn.svm library, opting for the RBF kernel. This approach, while effective, comes with challenges such as increased sparsity in high-dimensional spaces and sensitivity to feature selection. The computational demands of SVMs grow with the number of features, leading to longer training times and higher memory consumption.

We employed various machine learning algorithms for data classification. The evaluation metrics for assessing the model include accuracy, balanced accuracy, Macro-F1, MCC, and training time. Through comprehensive comparison and evaluation, we determined that Random Forests exhibit superior performance. The performance comparison among different algorithms is presented in Table 8.

5.3. The Final Pipeline Performance Evaluation

Finally, we conducted experiments using the Random Forest algorithm with the hyperparameters $n\_estimators = 150$ and $max\_depth = 25$, while utilizing nine feature parameters for classification. The mixed data were employed as the final training set. The accuracy achieved on the mixed data test set is 95.0%, while the accuracy on the LEIA test set is 97.8%. The specific details of the final test results are presented in Table 9. Figures 10 and 11 display the confusion matrices for the LEIA test set and the mixed data test set.

In the mixed data test set, the classification results remain unsatisfactory for certain classes, such as HLGRB, SN_SBO,

Table 9
Performance Evaluation of the Final Pipeline

Data	Accuracy	Balanced Accuracy	Macro-F1	MCC
Mixed data test set	95.0%	89.2%	85.4%	92.6%
LEIA data test set	97.8%	95.5%	94.3%	95.5%

and others, which comprise a small number of rare time-domain targets. This is primarily due to the limited number of objects in these classes. The LEIA survey has not yet identified these rare time-domain objects, and the data used continue to be simulated data from the ROSAT star catalog. In contrast, cosmic ray targets lack a light curve, allowing for their successful identification based on other characteristic features. The accuracy rate for identifying cosmic ray targets reaches 100%.

6. Discussion and Application

Random Forests construct each decision tree by randomly selecting a subset of features, which effectively mitigates the risk of overfitting in high-dimensional spaces. Each tree is trained on a distinct subset of features, thereby reducing the model's dependence on any individual feature. Such diversity significantly enhances the model's generalization ability. Selecting appropriate features for modeling in high-dimensional spaces can be challenging, but Random Forests excel at handling numerous features without requiring explicit feature selection. The model can automatically identify significant features from the entire set and maintain relatively high performance, even in the presence of irrelevant features. Figure 12 illustrates the ranking of feature importance, highlighting the top nine features of significant importance.

6.1. Interpretability and Feature Importance

Figure 5 demonstrates the innovative aspects of our feature design. Extracting features from the light curve data in previous studies, such as power-law fitting and Lomb–Scargle periodogram, has proven highly challenging. Three distinct features within the light curve—"kurt," "skew," and "modulation index"–were identified as significant features. These three

Figure 10. Confusion Matrix on Mixed Data Test Set.

attributes capture the sharpness, skewness, and oscillatory behavior manifested in the light curve, respectively. The classification of cosmic rays and stars relies significantly on these light curve features. Cosmic rays have almost no light curve characteristics, only photons concentrated in a single readout frame. Stars have weak variability, occasionally exhibiting a stellar flare.

Regarding spectral features, we opted not to utilize the hardness ratio method, which may not capture all the characteristics of the X-ray spectrum. Instead, we enhanced previous research by replacing the hardness ratio with energy band counting, leading to improved outcomes. Our study revealed that features associated with energy spectra distribution, specifically "b_hard" and "c_hard," exhibited remarkable significance in the model. This can be attributed to the primary role of EP as an X-ray telescope, specializing in the study of high-energy celestial objects. Three distinct features are capable of capturing the energy spectral distribution across the soft to hard X-ray energy bands for different classes. Therefore, incorporating energy spectral distributions as features provides substantial advantages in this specific context. These features make significant contributions by providing valuable information and aiding in the classification process.

For instance, "b_hard" and "de_hard" made valuable contributions to classifying SNR, XRB, and cluster of galaxies, effectively distinguishing or excluding these classes and thereby improving the screening of AGNs, among others.

Additionally, utilizing "galactic longitude" and "galactic latitude" provides an efficient means of classifying galactic sources, including XRBs and pulsars. Table 10 describes the high contribution features for different classes in LEIA data.

Simulated data frequently include repeated observations of the same celestial sources, resulting in the presence of multiple observations for individual sources. We recognized that directly utilizing spatial information could lead to excessive overfitting of sources with multiple observations within the spatial feature space. To address this concern, we employed a resampling technique, specifically SMOTE, on the "galactic longitude" and "galactic latitude." Consequently, the spatial positions of the sources displayed a randomized distribution across the sky map, as illustrated in Figure 13.

6.2. Comparison of Different Methods for Constructing the Training Set

We assessed the effectiveness of three data set construction methods: utilizing only EP simulation data, only LEIA data,
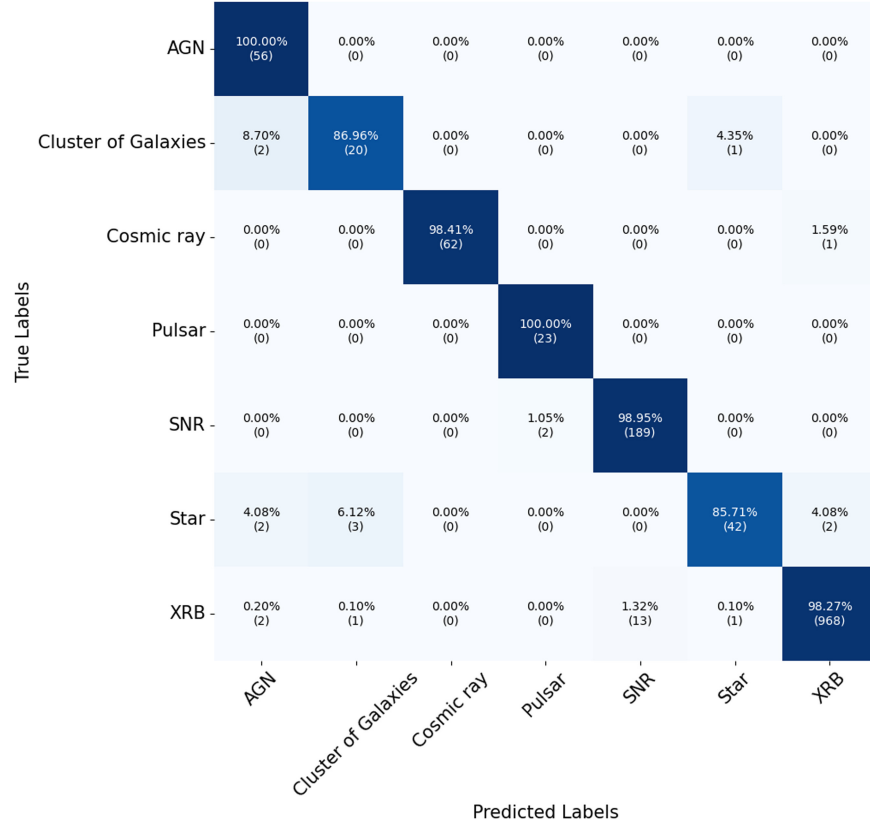
Figure 11. Confusion Matrix on the LEIA Data Test Set.

and merging both EP and LEIA data. The evaluation was conducted using a consistent set of features. Given the model's intended use as a pipeline for classification tasks in both LEIA and EP, we measured the efficacy of these three training set construction methods using untrained LEIA data as the test set.

In the instance where simulated data served as the direct training set, SMOTE resampling was applied uniformly to each class to achieve 16,000 data points per class and a Random Forest classifier was utilized. Notably, the simulated data lacked certain categories detected by LEIA, such as SNR, cosmic rays, and clusters of galaxies. Consequently, the categories common to both data sets were restricted to XRB, AGN, and stars, exclusively reserved for testing in the LEIA test set.

When employing actual LEIA observational data as the training set, the SMOTE resampling strategy was adjusted to accommodate the smaller data set size, resulting in 3387 samples per class. However, the test set featured fewer categories due to the unique presence of classes in the simulated data not observed by LEIA. Classification testing was carried out using a consistent approach. It is essential to highlight that in scenarios with limited data volume, there is a

risk of overfitting post-resampling. Detailed outcomes for these three scenarios are outlined in Table 11.

Given the current operational timeframe, data scarcity, and observation region constraints, it is crucial to acknowledge these limitations despite the promising results derived from utilizing LEIA data for training. Additionally, LEIA data cover a narrower spectrum of source categories. Notably, anticipated future surveys by both LEIA and EP are projected to detect specific time-domain targets like SGRB and SN_SBO, present in simulated data but not yet observed. Classifying these targets will aid in identifying novel celestial objects. Thus, we have opted to continue incorporating EP simulated data alongside LEIA data in our training model.

6.3. Pipeline Application

The trained classification model has been encapsulated in a Docker container and integrated into the data processing pipeline. The classification model delivers AI-based classification outcomes for each observation, along with the corresponding probability of the predicted class. The processing time for a single observation within the pipeline is approximately 0.19 s, while an observation may encompass multiple sources.
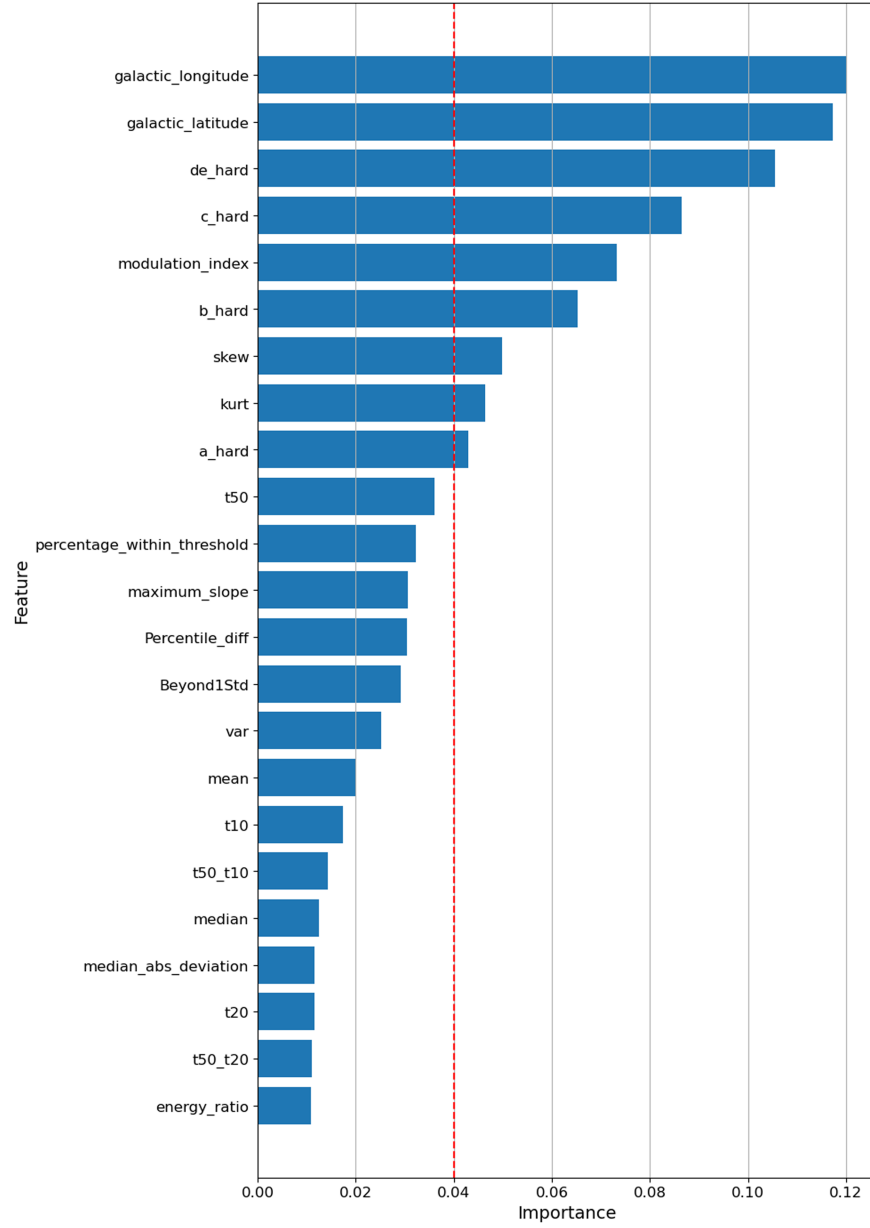
Figure 12. Feature Importance ranking.

Table 10
High Contribution Features for Different Classes in LEIA Data

Class	High Contribution Features
AGN	a_hard, kurt, galactic longitude, galactic latitude
XRB	galactic longitude, galactic latitude, de_hard, c_hard
Cluster of Galaxies	galactic longitude, galactic latitude, a_hard, b_hard
Cosmic Ray	kurt, skew, modulation index
Pulsar	galactic longitude, galactic latitude, b_hard, c_hard
SNR	galactic longitude, galactic latitude, b_hard
Star	kurt, skew, galactic latitude, c_hard

This capability significantly aids the Transient Advocate team in validating observed sources. Figure 14 showcases the automated classification of observation sources within the EP data processing interface. Actual observation data from LEIA in October and November were selected to evaluate the classification results of the application model. After filtering out interference items like arm and fake sources from this data set, a total of 596 instances were analyzed. The classification accuracy for these data stands at 86.7%, with the classification confusion matrix depicted in Figure 15.
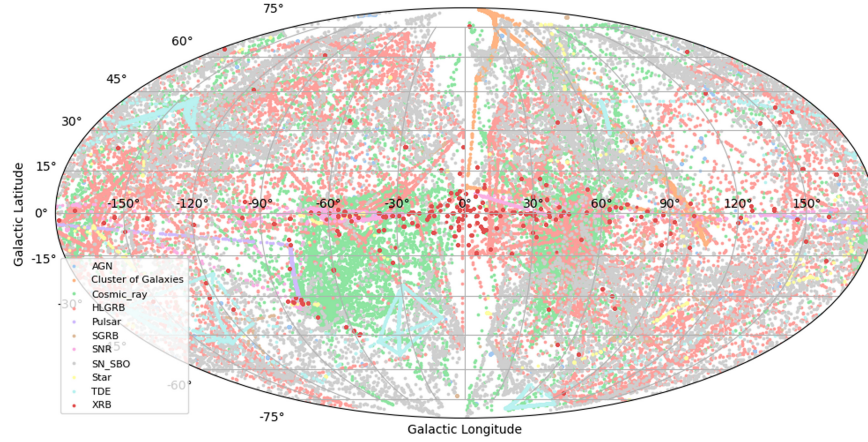
Figure 13. The distribution of the data on the sky map after SMOTE resampling.

	Source ID	Version	Obs Time (UT)	RA(J2000)	Dec(J2000)	Pos_E	Esti_Flux	Ref Flux		Type	Classification	AI_Classification	AI_Prob
										All			
1	ep06800007190wxt15s1	v1	2023-11-10 17:12:08	80.207	-71.92	0.519	3.35e-10	1.66e-10	🔒	known_source	LMXB	XRB (LMXB, HMXB, Cataclysmic Binary, X-ray binary)	0.986666666666667
2	ep06800007190wxt15s2	v1	2023-11-10 17:12:08	84.739	-64.094	0.555	3.03e-10	4.47e-10	🔒	known_source	HMXB	XRB (LMXB, HMXB, Cataclysmic Binary, X-ray binary)	0.693333333333333
3	ep06800007190wxt15s3	v1	2023-11-10 17:12:08	81.328	-69.628	0.737	1.66e-10	4.83e-10	🔒	known_source	SNR	SNR	1.0
4	ep06800007190wxt15s4	v1	2023-11-10 17:12:08	84.917	-69.731	0.718	1.82e-10	1.11e-9	🔒	known_source	HMXB	XRB (LMXB, HMXB, Cataclysmic Binary, X-ray binary)	0.92
5	ep06800007190wxt15s5	v1	2023-11-10 17:12:08	83.946	-66.047	0.979	1.11e-10	7.11e-11	🔒	known_source	SNR	SNR	1.0
6	ep06800007190wxt15s6	v1	2023-11-10 17:12:08	82.171	-65.445	0.754	4e-11	4.93e-11	🔒	known_source	T Tauri Star	star	0.786666666666667
7	ep06800007190wxt16s1	v1	2023-11-10 17:12:08	71.809	-65.864	0.959	4.83e-11		🔒	known_source	Nova / Star	SNR	0.633333333333333
8	ep06800007191wxt15s1	v1	2023-11-10 18:46:08	80.174	-71.927	0.625	2.39e-10	1.66e-10	🔒	known_source	LMXB	XRB (LMXB, HMXB, Cataclysmic Binary, X-ray binary)	0.973333333333333
9	ep06800007191wxt15s2	v1	2023-11-10 18:46:08	84.748	-64.088	0.516	2.55e-10	4.47e-10	🔒	known_source	HMXB	XRB (LMXB, HMXB, Cataclysmic Binary, X-ray binary)	0.566666666666667
10	ep06800007191wxt15s3	v1	2023-11-10 18:46:08	84.935	-69.753	0.725	1.79e-10	1.11e-9	🔒	known_source	HMXB	XRB (LMXB, HMXB, Cataclysmic Binary, X-ray binary)	0.953333333333333
11	ep06800007191wxt15s4	v1	2023-11-10 18:46:08	81.312	-69.63	0.768	1.49e-10	4.83e-10	🔒	known_source	SNR	SNR	0.993333333333333
12	ep06800007191wxt15s5	v1	2023-11-10 18:46:08	83.946	-66.039	0.958	9.32e-11	7.11e-11	🔒	known_source	SNR	SNR	0.993333333333333
13	ep06800007191wxt16s1	v1	2023-11-10 18:46:08	71.796	-65.864	0.902	8.08e-11		🔒	known_source	Nova / Star	SNR	0.673333333333333
14	ep06800007192wxt15s1	v1	2023-11-10 20:20:08	14.159	60.719	0.596	1.3e-10	1.04e-10	🔒	known_source	Be Star	SNR	0.46

Figure 14. Automatic classification of observation sources in the EP data processing interface.

Table 11
Comparison of the Effects of Three Methods of Constructing Training Sets

Training Set	Accuracy	Balanced Accuracy	Macro-f1	MCC
Only Simulate data	90.8%	28.5%	47.4%	36.3%
Only LEIA data	98.2%	97.0%	96.0%	96.4%
Mixed data	97.8%	95.5%	94.3%	95.5%

Upon examination, it was noted that misclassifications within the AGN category were prevalent across multiple observations from three sources, primarily being classified as stars. AGNs and stars share similar physical characteristics, leading to potential confusion. In the case of pulsars, misclassifications were observed in multiple observations from two sources. One of them, the magnetar, was misclassified as a star or an SNR, largely due to physical resemblance. The other source, a pulsar, was classified as an XRB, primarily influenced by its proximity to the Galactic center. The galactic location erroneously attributed to XRB was based on feature

contributions. Lastly, the classification performance for clusters of galaxies is suboptimal due to the limited representation of galaxy clusters in the training set, impacting the model's overall performance.

6.4. Limitation

The data set includes a limited number of instances belonging to the cluster of galaxies and HLGRB. Consequently, these minority class instances may face challenges during SMOTE resampling, increasing the risk of overfitting. This risk arises from the potential generation of synthetic samples that closely mimic the minority class instances, potentially exacerbating overfitting concerns. Additionally, the application of SMOTE can lead to increased class overlap, blurring the distinctions between classes and impacting the model's decision boundary. This could make it more difficult to differentiate between different categories. For example, the resampling process had a notable impact on the positional distribution of the cluster of galaxies, as illustrated in
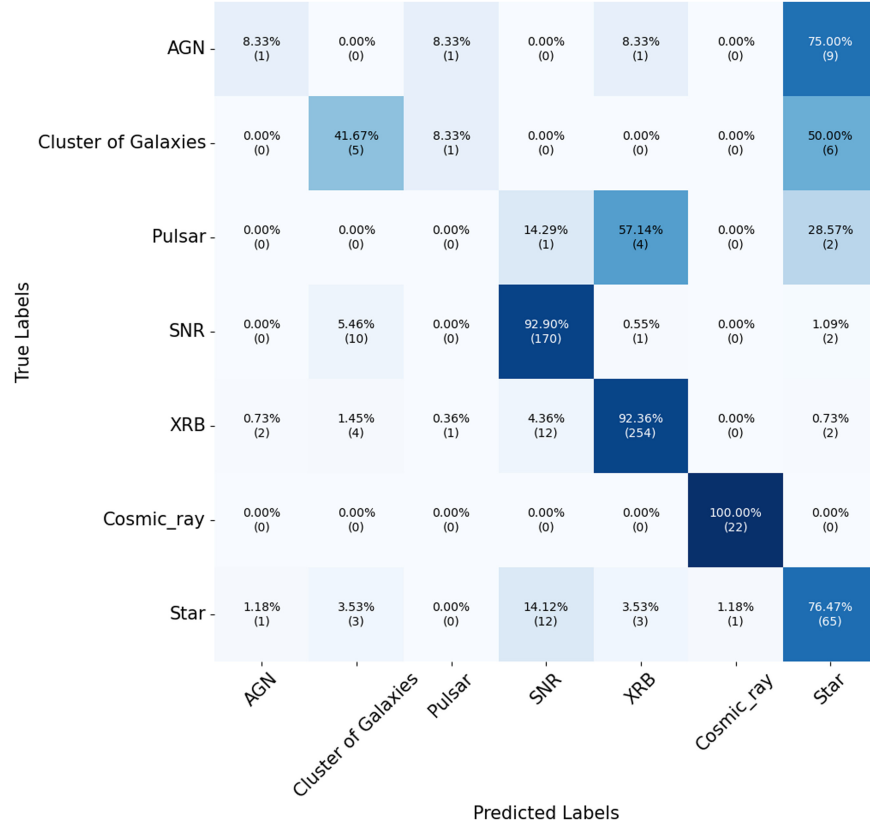
Figure 15. Confusion Matrix for Classification of Models in LIEA Observational Data in 2023 October and November.

Figure 13. The availability of more extensive data in the future is expected to alleviate this limitation.

The features of "galactic longitude" and "galactic latitude" are indicative of the distribution of sources on the sky map. They are particularly effective in classifying sources that are in the Galactic center. However, when dealing with sources that are situated at significant distances from the galactic center and possess high galactic latitudes, such as high-latitude XRBs and SNRs, these features may exert an inverse effect within the classifier, potentially leading to misclassification.

The subpar classification performance also can be attributed to frequent observations from diverse sources, leading to inadvertent classification errors. Notably, a significant portion of misclassifications exhibit a classification probability below 0.5, typically hovering around 0.3 or 0.4. To mitigate this issue, we plan to introduce a probability threshold as a filter in our forthcoming work to enhance the classification accuracy.

7. Conclusion

The paper primarily delves into investigating a time-domain target classification algorithm tailored for X-ray telescopes. This research is executed through empirical analysis utilizing simulated data from EP and observational data from LEIA. A data set is curated by combining EP simulation data with LEIA measurements, and a distinct set of classification features tailored for X-ray telescope data is proposed. This approach showcases promising performance in scenarios characterized by limited data points and shorter observation durations. Following a comparative analysis of various machine learning algorithms, Random Forest is selected as the classification algorithm, achieving an accuracy rate of 97.8%. Moreover, this study integrates classification models into data processing pipelines, facilitating classification predictions for newly detected sources. The implications of this research are notably significant for the data processing tasks associated with EP missions. Upon EP's acquisition of fresh data, the classification model will be leveraged to classify categories not previously observed by LEIA. The findings presented in this paper can serve as a valuable resource for data analysis in high-energy space satellite missions and time-domain astronomy.

Appendix A
The Energy Spectrum for Each Class

This appendix presents figures of energy spectrum for each class described in Section 3. These classes include AGNs A1, XRBs A2, clusters of galaxies A3, cosmic rays A4, pulsars A5, SNRs A6, and stars A7 from LEIA data, as well as TDEs A8, SN_SBOs A9, and HLGRB A10 from EP simulation data.
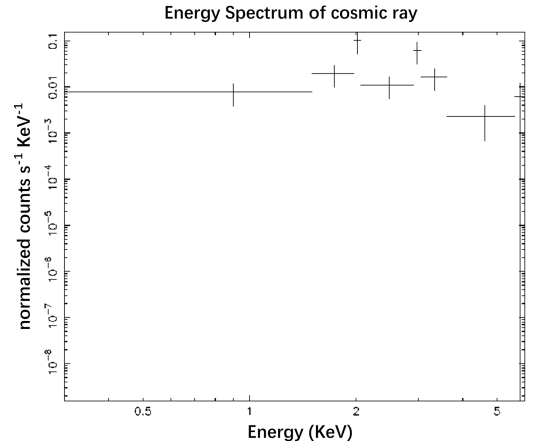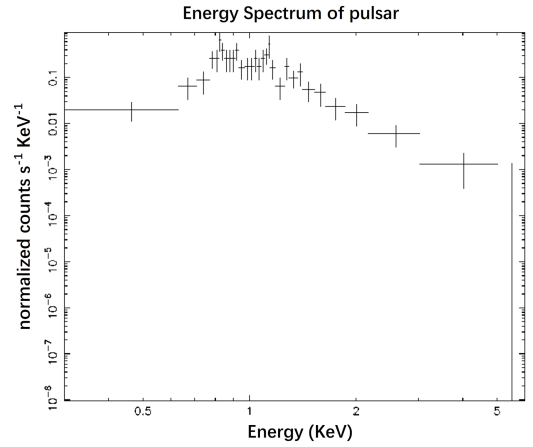

Figure A3. Energy spectrum of cluster of galaxies.

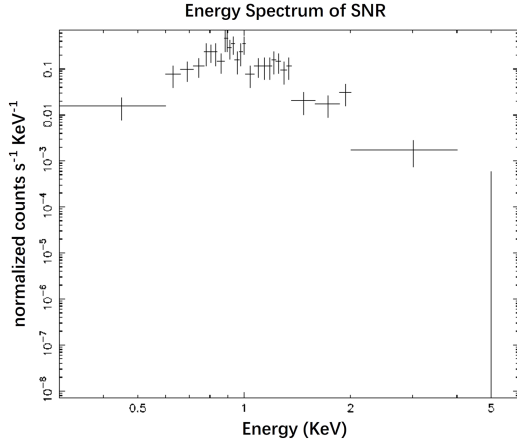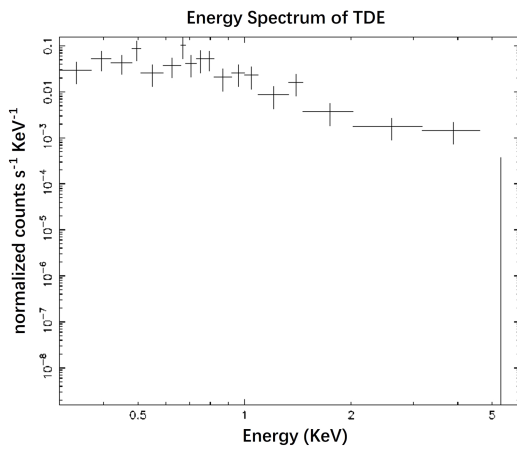**Figure A1.** Energy spectrum of AGN.

Figure A4. Energy spectrum of cosmic ray.

Figure A2. Energy spectrum of XRB.

Figure A5. Energy spectrum of pulsar.

Figure A6. Energy spectrum of SNR.
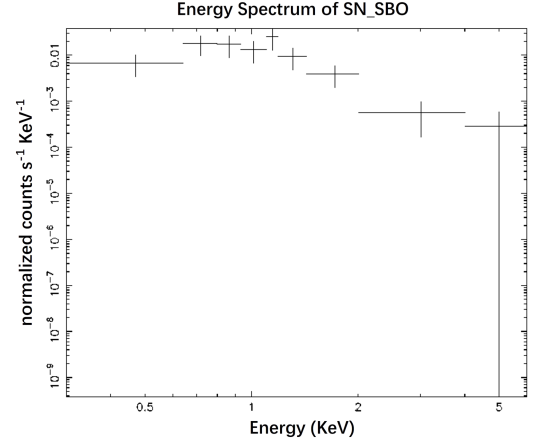
Figure A9. Energy spectrum of SN_SBO.
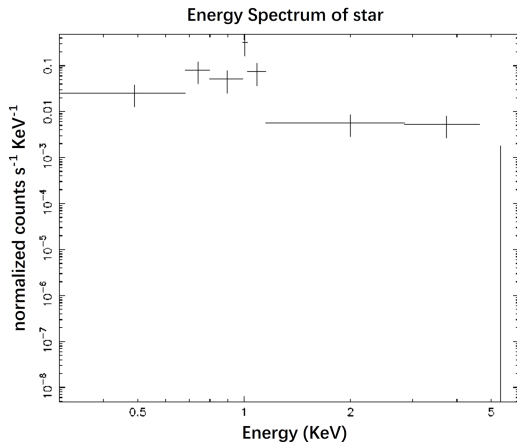
Figure A7. Energy spectrum of star.
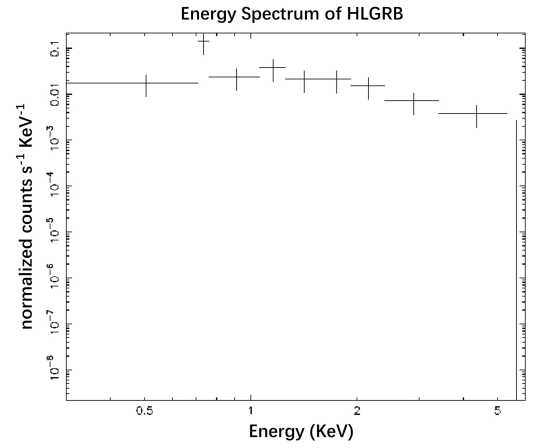
Figure A10. Energy spectrum of HLGRB.

Appendix B
The Light Curve for Each Class

This appendix presents figures of light curve for each class described in Section 3. These classes include AGNs B1, XRBs B2, clusters of galaxies B3, cosmic rays B4, pulsars B5, SNRs

Figure A8. Energy spectrum of TDE.
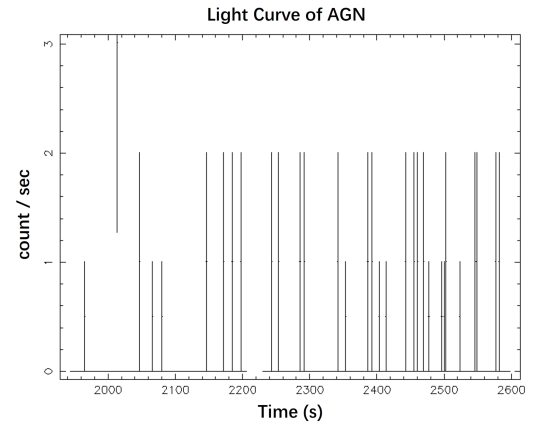
Figure B1. Light curve of AGN.
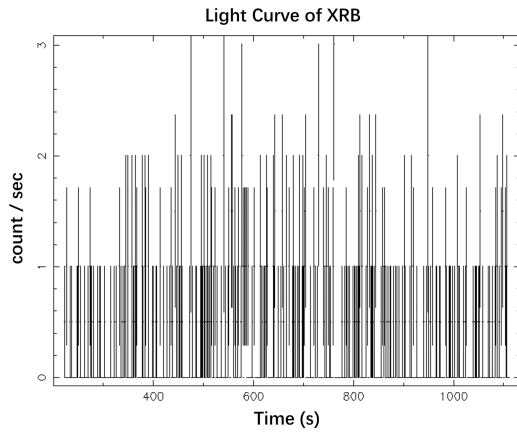
Figure B2. Light curve of XRB.
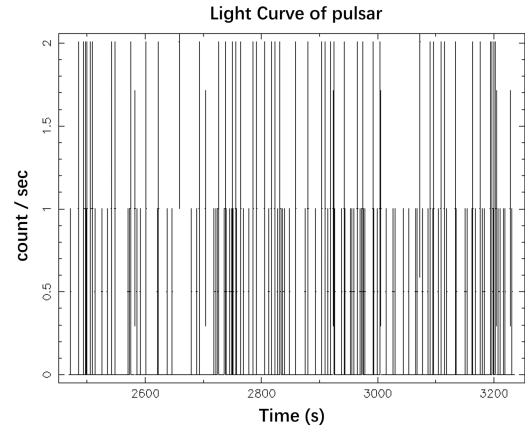
Figure B5. Light curve of pulsar.
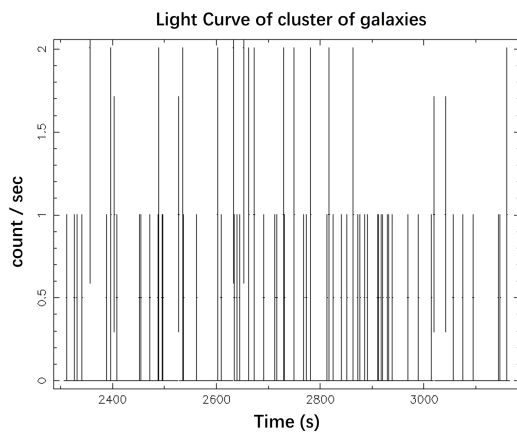
Figure B3. Light curve of cluster of galaxies.
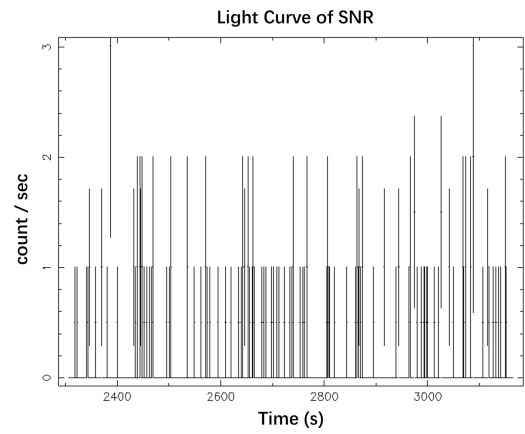
Figure B6. Light curve of SNR.
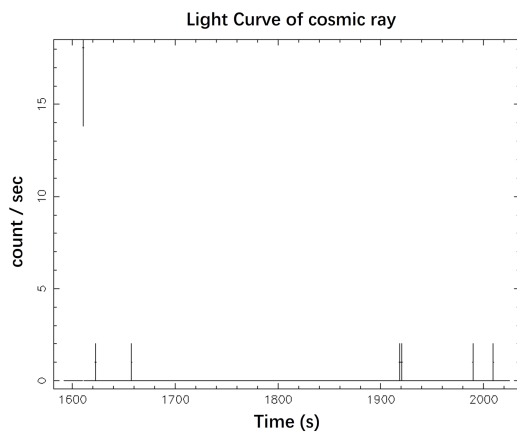
Figure B4. Light curve of cosmic ray.
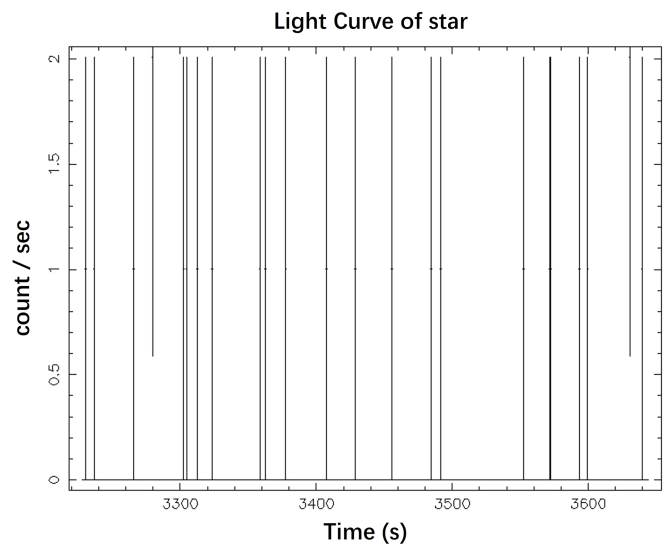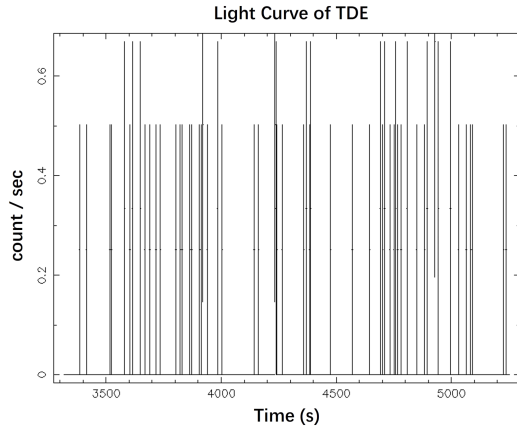
Figure B7. Light curve of star.
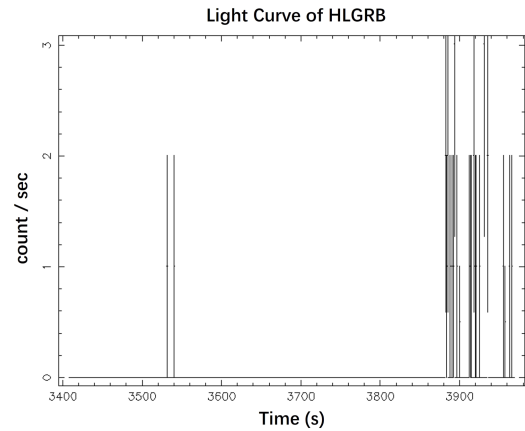
Figure B8. Light curve of TDE.
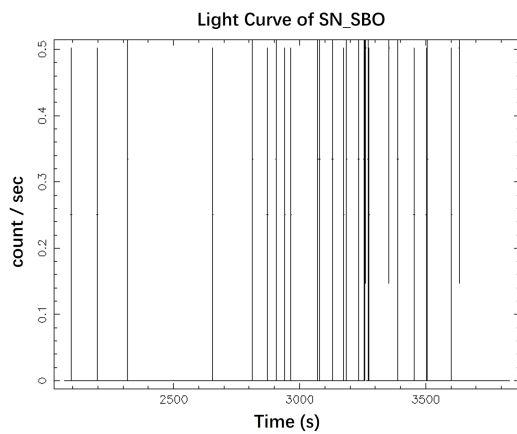

Figure B10. Light curve of HLGRB.

Figure B9. Light curve of SN_SBO.

B6, and stars B7 from LEIA data, as well as TDEs B8, SN_SBOs B9, and HLGRBs B10 from EP simulation data.

ORCID iDs

Xiaoxiong Zuo ⓘ https://orcid.org/0000-0002-0772-6280
Yihan Tao ⓘ https://orcid.org/0000-0002-3143-9337

References

Angel, J. 1979, Proc. SPIE, 184, 84
Breiman, L. 2001, Mach. Learn., 45, 5
Burrows, D. N., Hill, J., Nousek, J., et al. 2005, SSRv, 120, 165

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. 2002, J. Artif. Intell. Res., 16, 321
Chen, T., & Guestrin, C. 2016, Proc. of the 22nd ACM Sigkdd Int. Conf. on Knowledge Discovery and Data Mining (New York: ACM), 785
Cortes, C., & Vapnik, V. 1995, Mach. Learn., 20, 273
Cover, T., & Hart, P. 1967, ITIT, 13, 21
Jansen, F., Lumb, D., Altieri, B., et al. 2001, A&A, 365, L1
Jia, P., Liu, W., Liu, Y., & Pan, H. 2023, ApJS, 264, 43
Li, D., Ling, Z., Liu, Y., et al. 2022, ATel, 15754, 1
Ling, Z., Liu, Y., Zhang, C., et al. 2022, ATel, 15748, 1
Lo, K. K., Farrell, S., Murphy, T., & Gaensler, B. 2014, ApJ, 786, 20
McGlynn, T., Suchkov, A., Winter, E., et al. 2004, ApJ, 616, 1284
Murthy, S. K., Kasif, S., & Salzberg, S. 1994, J. Artif. Intell. Res., 2, 1
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, The Journal of Machine Learning Research, 12, 2825
René, H. 2010, XROI, 2010, 139148
Richards, J. W., Starr, D. L., Butler, N. R., et al. 2011, ApJ, 733, 10
Sun, H., Wang, C., Hu, J., et al. 2023, ATel, 16195, 1
Tranin, H., Godet, O., Webb, N., & Primorac, D. 2022, A&A, 657, A138
Virtanen, P., Gommers, R., Oliphant, T. E., et al. 2020, NatMe, 17, 261
Weisskopf, M. C., Tananbaum, H. D., Van Speybroeck, L. P., & O'Dell, S. L. 2000, Proc. SPIE, 4012, 2
Voges, W., Aschenbach, B., Boller, T., et al. 1999, A&A, 349, 389–405
Yang, H., Hare, J., Kargaltsev, O., et al. 2022, ApJ, 941, 104
Yang, H., Zhang, W., Jin, C., et al. 2023, ATel, 16352, 1
Yuan, W., Zhang, C., Feng, H., et al. 2015, arXiv, arXiv:1506.07735
Yuan, W., Zhang, C., Ling, Z., et al. 2018a, Proc. SPIE, 10699, 1069925
Yuan, W., Zhang, C., Ling, Z., et al. 2018b, SSPMA, 48, 039502
Zhang, C., Ling, Z., Sun, X., et al. 2022a, ApJL, 941, L2
Zhang, J., Qi, L., Yang, Y., et al. 2022b, APh, 137, 102668
Zhang, Y., Zhao, Y., & Wu, X.-B. 2021, MNRAS, 503, 5263