Highlights of Spanish Astrophysics XII, Proceedings of the XVI Scientific Meeting of the Spanish Astronomical Society held on July 15 - 19, 2024, in Granada, Spain. M. Manteiga, F. González Galindo, A. Labiano Ortega, M. Martínez González, N. Rea, M. Romero Gómez, A. Ulla Miguel, G. Yepes, C. Rodríguez López, A. Gómez García and C. Dafonte (eds.), 2025

Machine Learning classification of X-ray binaries from timing analysis: preparation for the THESEUS mission.

Nespoli, E.[1], Pérez-Suay, A.[2], Blay, P.[1], Suso, J.[3], Fabregat, J.[3], Gasent-Blesa, J.L.[2], Navarro-González, J.[2], Pascual-Venteo, A.B.[2]

[1] Universidad Internacional de Valencia (VIU), c/ Pintor Sorolla 21, 46002 Valencia, Spain
[2] Image Processing Laboratory, Universitat de València, c/ Catedrático José Beltrán 2, 46980 Paterna, Spain
[3] Observatorio Astronómico, Universitat de València, c/ Catedrático José Beltrán 2, 46980 Paterna, Spain

Abstract

The "Transient High Energy Sky and Early Universe Surveyor" (THESEUS) is a medium-sized space mission of the European Space Agency. Its scientific goal will be to improve the understanding of high-energy transient phenomena across cosmic time, including gamma-ray bursts, electromagnetic counterparts to gravitational wave or neutrino sources, magnetars, novae, X-ray binaries, and more. This work aims at identifying and characterizing high-energy transient objects, with particular emphasis on X-ray binary systems, employing Machine Learning techniques as part of the preparatory work for the scientific analysis of the THESEUS mission data.

1 Introduction

High-energy transients, such as gamma-ray bursts, X-ray binaries (XRBs), and supernovae, offer a window into some of the most energetic processes in the universe. The "Transient High-Energy Sky and Early Universe Surveyor" (THESEUS) mission [1], part of the European Space Agency's Cosmic Vision program, was selected for phase A study in November 2023, which will last until 2026, with a possible launch in 2037, pending approval. The mission will offer unprecedented observational capabilities in the X-ray and gamma-ray regimes, providing data that will improve our knowledge and understanding of these astrophysical objects.

XRBs exhibit variable X-ray emission over various timescales, ranging from milliseconds to years, depending on the mass transfer rate, accretion processes, and the nature of the compact object. The features observed in the light curves of these systems, along with their frequency-domain characteristics, contain valuable information about the XRB and the mass

exchange process, enabling classification of these systems based on their timing properties.

Given the high volume of data expected from THESEUS, manual classification of XRBs and other transient phenomena will not be feasible. Machine Learning (ML) approaches are therefore essential for automating the detection, classification, and characterization of XRBs and other transients in real time. The application of ML techniques, a relatively recent analytical tool in astrophysics, has shown significant effectiveness in classifying X-ray sources [3, 2, 11]. However, previous studies reveal a notable gap in the systematic analysis of source variability, highlighting a promising area for further research and methodological advances.

The automatic classification of XRBs is ideal for ML techniques for several reasons. The extensive data sets collected over many years of XRB observations are ideally suited for training ML algorithms, enabling better identification of subtle patterns and improving classification of these complex systems. XRBs exhibit complex physical processes like accretion, flaring, periodic outbursts leading to variable and noisy data. ML, especially deep learning, can automatically learn from such data, identifying hidden patterns in noisy observations. Relationships between observed features (e.g., flux, spectral states and light curves) and their underlying physical properties are in many cases non-linear. ML models like neural networks or random forest can effectively learn these non-linear dependencies. Additionally, ML models can adapt to diverse types of observational data, dealing with variations in observation conditions, sampling frequency or instruments sensitivity, remaining robust across a wide range of conditions, which supports a more generalizable understanding of X-ray sources.

As a first approach, in the work presented here we limited the study to a sub-class of XRBs, namely high-mass X-ray binaries (HMXBs), in which a compact object, either a neutron star (NS) or black hole (BH), accretes material from a massive stellar companion [4, 8]. These systems are characterized by intense X-ray emission from accretion, primarily driven by the massive companion's stellar wind, by the Be star's circumstellar disk or Roche-lobe overflow.

2 Data sets and methodology

2.1 Dataset description

Using long-term archival data from XMM-Newton [7] and MAXI [10], we tested a range of ML algorithms to evaluate the effectiveness of different feature extraction techniques. This work is a preparatory step for THESEUS and seeks to optimize ML methods that can handle the mission's large, complex datasets. Both XMM-Newton and MAXI operate at a similar energy range (0.5 - 12 keV), but their data have different sampling frequency, below 1s in the first case and of the order of 90 minutes in the second case. The different time resolution will eventually enable us to study short-term variability and medium-term variability and extract useful classification features from both modes, although this objective will be part of a subsequent, more complete work. Both science archives contain a wealth of observations of HMXBs and other high-energy astrophysical objects. We selected a subset of well-known HMXBs from the archives, ensuring diversity in the types of binaries included (BH-binaries and NS-binaries with either a Be or a supergiant companion). For those systems, we retrieved light curves from the archives and obtained Lomb-Scargle periodograms from them. By using

the frequency distribution given by these periodograms, Power Spectral Density (PSD) were then computed. Light curves, on one side, and PSDs on the other, were employed as input to ML techniques, with a total of 237 samples, divided into three different classes: 15 light curves and PSDs from Roche-lobe overflow accreting system, 93 from stellar wind accreting systems, and 129 from systems accreting from a Be star companion. We split the data into train/test sets, 66% and 33%, respectively.

2.2 Methodology

2.2.1 Machine Learning models

To classify the available light curves and PSDs, we used a set of five statistical models: three parametric and two non-parametric. All five methods are supervised learning techniques, utilising class label information during training. The three selected parametric models are based on linear decision rules. In particular, we choose the Linear Support Vector Machines (LinSVM) [13], which seeks the hyperplane that provides the largest margin between the nearest data points (called support vectors) of each class. The model minimises classification errors while maximising the margin. Also, we considered the Principal Component Regression (PCR)[9], a regression technique combining Principal Component Analysis (PCA) with linear regression. PCA reduces dataset dimensionality by transforming original variables into a smaller set of uncorrelated principal components that capture most of the data's variance, reducing models' complexity by focusing on the most significant components. Additionally, we employed Partial Least Squares (PLS) [14], which, like PCR, reduces the dimensionality of the data. In contrast to PCR, PLS finds components that maximize the covariance between the inputs and the target. PLS is especially useful when there are more dimensions than observations, which is the case in our data. We adapted the regression models to classification using one-hot encoding. The two non-parametric models consist of the k-nearest neighbors algorithm (KNN) [5] and the kernelised version of Support Vector Machines (KerSVM) [12]. KNN makes predictions by storing the entire training dataset and classifying new data points based on the majority class of the k-nearest neighbors, using distance metrics like Euclidean distance. In contrast, KerSVM is a model-based algorithm that finds an optimal hyperplane to separate data points of different classes with the largest possible margin. KernSVM in contrast to LinSVM uses a kernel function to transform the input data into a higher-dimensional space by means of the "kernel trick" [12], where it may become linearly separable. We employed the Radial Basis Function (RBF) kernel, which measures the similarity between two input vectors in the transformed feature space.

2.2.2 Light curves preprocessing

Data pre-processing is critical in time-series analysis, especially when dealing with long-term light curves from different high-energy missions, which are often sparse, noisy, and irregularly sampled. These challenges make feature extraction and dimensionality reduction essential components of the workflow. A solution to different length in time series is to warp them to a common length l. We compare three different ways to choose the length parameter l of the time series. The first approach, referred to as Short Raw, sets the target length l

to the length of the shortest time series in the dataset. Here, all time series are truncated to this length, without applying any interpolation. The second approach, termed Short Interpolation, truncates each time series to the shortest length in the dataset, followed by linear interpolation to align the time points across all series with consistent time bins. Finally, we used the linear time warping (LTW) methodology [6]. In this approach, the target length l is set to the maximum length among all time series. To match this length, each time series is extended by repeating specific time points, ensuring that every x_i is warped to the specified target length l.

3 Results

Figure 1 presents a comparative analysis of the classification rates achieved by five different ML models on light curves (first row) and PSDs (second row), averaged over 25 independent train/test splits. Light curves have been processed with three different methodologies as explained in Section 2.2.2. PSDs are directly classified by means of the five ML models as all of them have the same length. In terms of methodology performance of light curves classification, LTW consistently performs well across various light curve types with both mission datasets, demonstrating its effectiveness in classifying HMXBs.

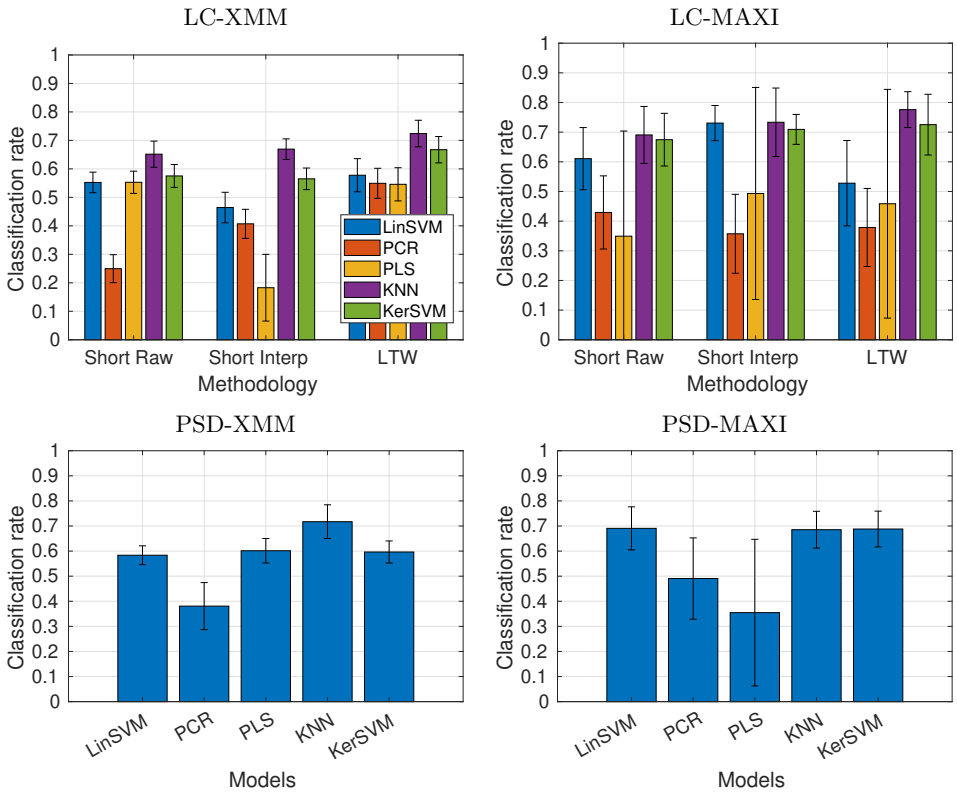

Figure 1: Averaged classification rate of light curves (1st row) and PSDs (2nd row).

The top-performing models for both light curves and PSDs are KNN and KerSVM, achieving around 0.7 in classification rate for all the used data. It is worth noting that the standard deviations are relatively small across all models (except PLS).

Figure 2 illustrates confusion matrices which measure the performance of KNN (first row) and KerSVM (second row), the two models which reported best results in classification rate. Each cell in a confusion matrix represents the number of instances that were actually of a certain class (true class) but were predicted as another class (predicted class). Both models exhibit similar behaviour, except for XMM light curves, where KNN shows more variability in prediction errors, and KerSVM mislabelled all Roche-lobe overflow cases, possibly because they are under-represented. Conducting an in-depth analysis of misclassification errors could provide insights into the causes of model failures.

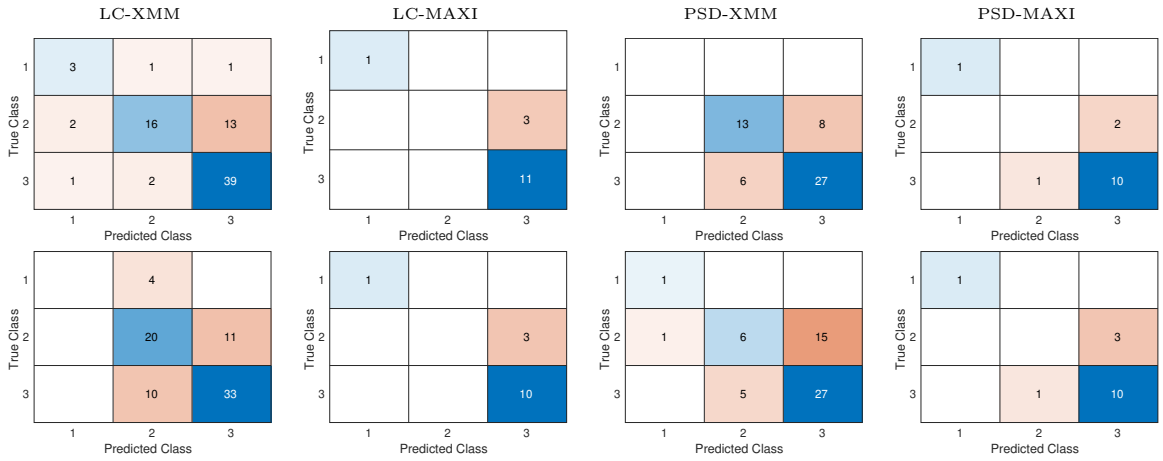

Figure 2: Averaged confusion matrices for light curves and PSDs classification. First row for KNN, second row for KerSVM. Classes refer to the type of mass-transfer during accretion, where (1) corresponds to Roche-lobe overflow, (2) to stellar-wind accretion, and (3) to accretion from a Be-star companion.

4 Conclusions and future work

Within the science goals of the THESEUS mission, which will be dedicated to improving the knowledge of high-energy transient phenomena across time, the study of XRBs is a key topic.

The application of ML techniques offers significant advantages in the identification and classification of Galactic XRBs, especially in contexts requiring the analysis of large datasets. This preliminary work has focused specifically on HMXBs to explore the potential of these approaches.

In this study, five ML techniques were applied to light curves and PSDs obtained from the XMM-Newton and MAXI mission archives. Initial results are encouraging: light curves appear to yield more accurate classifications than PSDs, with the KNN algorithm showing

the best performance. The selected models perform effectively with sparse-sampling. These results however will require further validation with additional datasets and analysis to confirm their robustness and effectiveness. It will be beneficial to explore methods designed to handle unbalanced classes, as the current dataset imbalance impacts the behaviour and performance of the ML models. Another possibility to palliate this effect is by increasing the number of samples in the minority classes. This can be achieved by utilising other mission archives or employing data augmentation techniques, such as adding noise to the observations.

Future efforts will focus on expanding the data and refining the models used in this study to enhance accuracy and broaden the scope of XRB analysis. Data will be extended to include Lomb-Scargle periodograms and additional datasets from missions such as SWIFT, RXTE/ASM, INTEGRAL/JEM-X, GINGA, and GRO. The aim is to ultimately apply this approach to all XRBs, including simulated THESEUS data to better anticipate future observations. On the modelling side, we will work to improve the existing models to maximise classification accuracy and apply neural network models to leverage their potential in complex data interpretation. Additionally, we will explore the feasibility of real-time identification of targets as soon as they are observed by THESEUS.

Acknowledgments

Grant PID2023-149817OB-C33 funded by MICIU/AEI/10.13039/501100011033, “ERDF A way of making Europe”, by “ERDF/EU” by the “European Union”. Project VIU24007 by ESenCIA/VIU.

References

[1] Amati, L., et al., 2021, Exp. Astron., 52, 183–218

[2] De Beurs, Z., et al., 2022, ApJ 933, 116

[3] De Luca, A., et al., 2021, A&A 650, A167

[4] Fornasini, F. M., et al., 2023, arXiv:2308.02645

[5] Fix, E., & Hodges, J.L., 1951, Discriminatory Analysis: Nonparametric Discrimination: Consistency Properties: USAF School of Aviation Medicine.

[6] Hiroshi, S., et al.: Advances in Neural Information Processing Systems: MIT Press, 2001.

[7] Jansen, F., et al., 2001, A&A, 365, L1-L6

[8] Kretshhmar et al., 2017, arXiv:1706.03969

[9] Mansfield, E. R., Webster, J. T., & Gunst, R. F., 1977, Journal of the Royal Statistical Society. Series C (Applied Statistics), 26, 34–40

[10] Matsuoka, M., et al., 2009, Publications of the Astronomical Society of Japan, 61, 999-1010

[11] Pérez-Díaz, V., et al., 2024, MNRAS 528

[12] Scholkopf, B, & Smola, A. J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond: MIT Press, 2001. - ISBN 0-262-19475-9

[13] Vapnik, V. N.: The nature of statistical learning theory : Springer-Verlag New York, Inc., 1995.

[14] Wold, S., Sjöström, M., & Eriksson, L., 2001, CILS, 58, 109–130