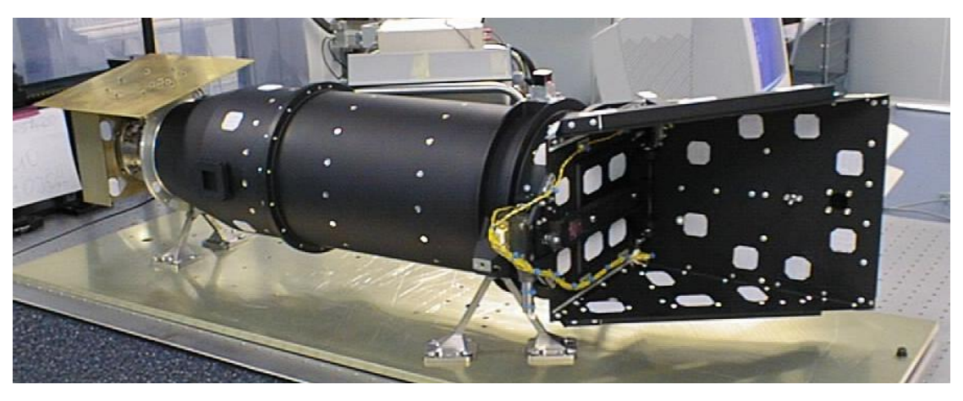
INTEGRAL-OMC Light Curve Classification with Transformers

John F. Aguilar[1,2]*, J. Alfonso−Garzón[3], Patricia Cruz[3] , A. Domingo[3], J.M. Mas−Hesse[3], Enrique Solano[3]

1. Departamento de Matemáticas, Universidad Militar Nueva Granada, Colombia.

2. PhD Programme in Astrophysics, Doctoral School, Universidad Autónoma de Madrid, Spain.

3. Centro de Astrobiología (CAB), CSIC-INTA, Madrid, Spain



We present the classification of variable stars from the first INTEGRAL-OMC catalog of optically variable sources (Alfonso-Garzón et al. 2012). This first version of the catalogue contains 5263 variable sources. For this work, we have adopted a subset of 1337 objects with calculated period and well-recognized classification in catalogs such as the International Variable Star Index (VSX). We adopted a method based on a Machine Learning transformer architecture to classify the OMC light curves, analyzing both the time series and the phase-folded light curves. This classification includes intrinsic (e.g., pulsating, eruptive) and extrinsic (e.g., eclipsing binaries, rotating stars) variable sources, and in the future will also include some of their subclasses, such as Algol, beta Lyr, W UMa, Cepheids, DSCT, LPVs, RR Lyrae (RRAB, RRC), RV Tauri, ACV, Ellipsoidal, Orion, T Tauri, Bes, and BYDracs. In this poster, we present our preliminary results of applying our algorithm to the subset described above. This methodology will be applied to the final INTEGRAL-OMC catalog with over 25,000 sources.

Introduction

INTEGRAL (INTErnational Gamma-Ray Astrophysics Laboratory) is a European Space Agency (ESA) mission launched in 2002 with the purpose of observing high-energy sources such as X-ray binaries, AGN, etc. This mission has several instruments, among which stands out the Optical Monitoring Camera (OMC, Mas-Hesse et al. 2003). This optical camera monitors the emission in the V-Johnson band (500-600 nm) from the sources observed by the X-ray and gamma-ray instruments, but it also obtains photometric measurements of many other sources in the field of view. OMC provides long-term light curves of thousands of potentially variable sources, and this fact motivates the preparation of algorithms to automate the classification of these sources. An important work in which this classification was implemented was in Alfonso-Garzon et al. 2012, where a study was made of a first catalog of object classification among which are: These classes include intrinsic variable sources (e.g. pulsating, eruptive) and extrinsic (e.g. eclipsing binaries, rotating stars) with objects with good photometric quality and more than 300 data points each. In most of the cases, the long time baseline of the OMC light curves provides us with good temporal coverage and well sampled phase-folded light curves (as it is shown for the object IOMC 1025000045 in Fig. 1, left). However, in other cases, these objects have import gaps between data points, such as IOMC4030000090 (Figure 1) that generate significant challenges for their classification and the estimation of their period, which in this case has been calculated with the phase dispersion minimization technique (PDM).

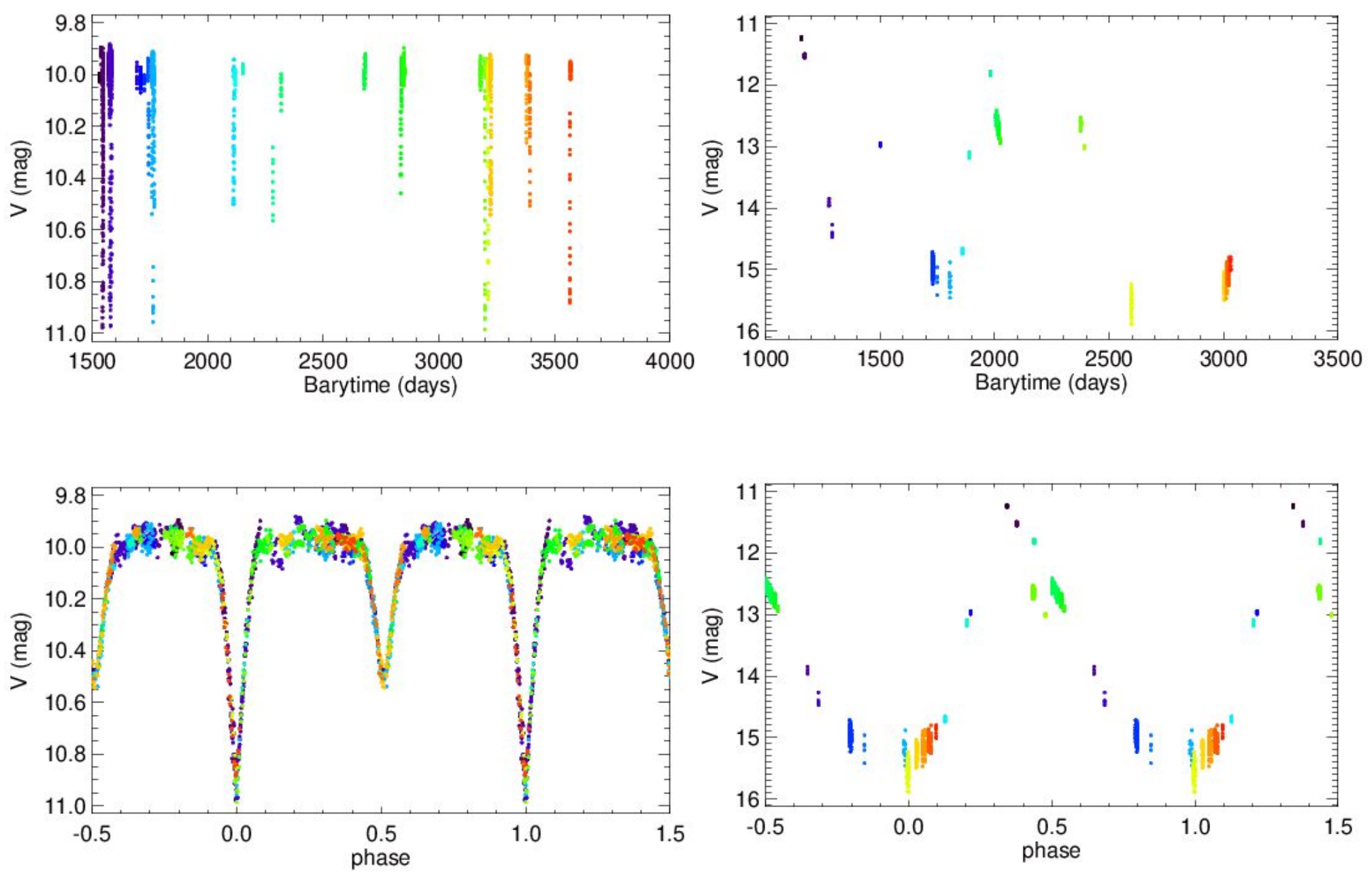


Figure 1. Top. Unfolded light curve of the variable objects IOMC 1025000045 (left) IOMC 4030000090 (right). Bottom. Phase-folded light curve of the same objects. Adapted from Alfonso-Garzon et al. 2012.

These data have been used as a first input for the preparation of algorithms that perform the automatic classification of these objects. Regarding the training algorithm for the classification of this type of data (Time Series), we will test the ML transformers architecture Ashish Vaswani et al. 2017, which promises simpler training and better scalability (Mario Morvan et al. 2022). The technique will be compared with others, such as the use of deep learning algorithms (such as DistClassiPy by Siddharth Chainiet al. 2024) to test its efficiency. Once the best approach for LC classification is found, the developed methodology will be applied to all OMC LCs and the results will be included in a future final OMC catalogue. The implementation of these objects will be implemented with phase-folded objects, which requires that they have the defined period, as well as their normalized curves, to highlight the shapes that characterize each of the classes under study.

Sample Selection

In order to generate a training data set, we have taken a sample of light curves (LCs) of variable stars from Alfonso-Garzon et al. (2012), which are representative of the future data that will compose the final OMC-INTEGRAL catalog of variable sources. Of the 5263 available objects, only those that had information about their period have been taken into consideration in the first stage, that is a subsample of 1337 variable sources, including eclipsing binaries (EB), eruptive stars (Er), pulsating stars (P), rotating stars (R), cataclysmic variables (CV), X-ray binaries (XB), extragalactic sources (AGN), and those just classified as variable stars (vstar). However, some of the mentioned categories (CV, XB, AGN, and vstar) are poorly represented in our initial sample, and have been excluded from this preliminary analysis, since with just a few identified sources of those types is not possible to train the algorithm. In Table 1, we can see the total number of sources considered in this analysis.

Type	Count
Eclipsing	559
Eruptive	21
Pulsating	663
Rotating	60

Table 1: List of objects taken from Alfonso-Garzon et al. 2012, which we will take into account for the implementation of the automatic classification algorithm

Methodology

Nowadays, there is a wide variety of time series classification algorithms that can be applied to the current study of light curves. However, transformer models have generated a revolution in what refers to the implementation of feature chains, especially in natural language processing (NLP), as positioned by Vaswani et al. (2017), generating important changes in the architectures of recurrent and convolutional networks, by focusing on attention mechanisms that model word dependencies in sequences, are trained to study the importance of each word in a sequence with respect to the others. Among the transformer-based architectures, we have employed BERT (Bidirectional Encoder Representations from Transformers, Ming-wei Chang Kenton et al. 2017.), which since its development has been used for text analysis and classification tasks, among other tasks. This model is based on the analysis of sentences taking into account previous and subsequent words and their relationship between them This characteristic and interpreting time series as sequential data on a time scale, allows capturing complex relationships in temporal data.

To adapt BERT for time series analysis, we preprocess the time series by converting them into a suitable format for BERT's tokenization process. Instead of working with words, as in natural language processing, we treat each value in the time series as a sequential token.

This step can involve discretizing continuous time series values or applying embeddings to represent each time point in the sequence. This process helps to highlight characteristics in our light curves despite the presence of different gaps between each of the objects, allowing the time series converted into tokenized time series to be classified by the BERT pre-trained model. Additionally, we apply positional encodings to ensure that the temporal order of the data is preserved, allowing BERT to capture dependencies both forward and backward in time. This enables the model to identify complex patterns in the time series despite irregular intervals or gaps between data points and classify them. This process is iterative and can be seen in Figure 2. This modification allows transformer-based models to handle temporal data, applying the same principle of contextual attention to capture temporal relationships. After tokenization, each value in the light curve can be associated with a vector that represents the relationship with other points in the series, allowing the algorithm to detect the main patterns in each of the selected classes, creating attention layers that capture both localized patterns and global trends across the time series. Considering that the amounts of data are unbalanced, a loss function called CrossEntropyLoss has been applied. This function measures the difference between the true labels of the time series and the predictions made by the model and validates that the model pays more attention to the less represented classes, which is important in unbalanced datasets.

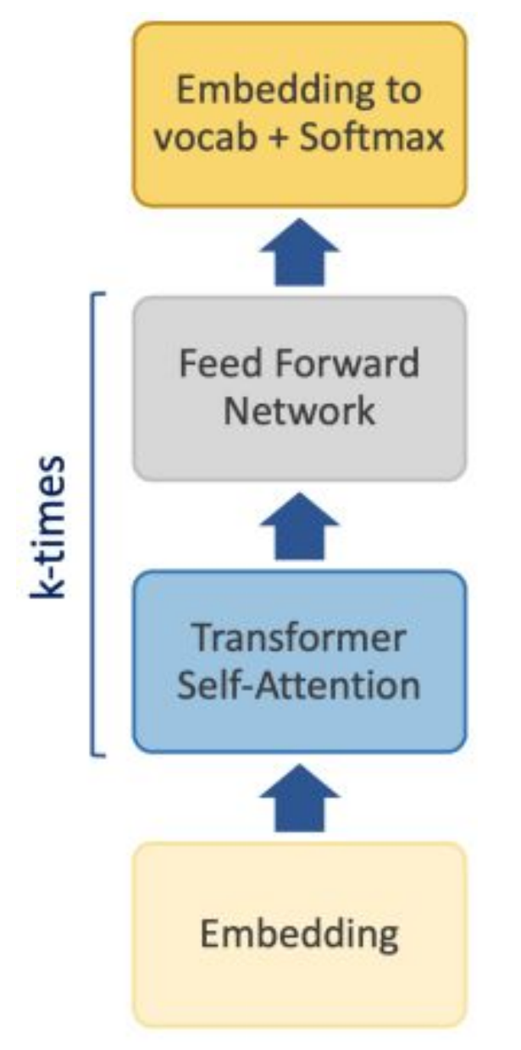


Figure 2: Description of the BERT model's feature capture process. This diagram illustrates the multi-layer structure involving self-attention and feed-forward networks. This graph is taken from Ozturk et al. 2017.

Results

When implementing the BERT architecture, 80 % of the available data (1042 objects) was taken as a training set and the remaining 20 % was taken as a test set to validate the trained model. This implementation resulted in a classification accuracy of 94%, as reported in the confusion matrix in figure 3.

When validating the number of objects that were not successfully classify it is observed that the proportion of misclassified objects was higher the Eruptive and Rotating object cases, which could be linked to the low number of objects in the study sample compared to the Eclipsing a Pulsating samples.

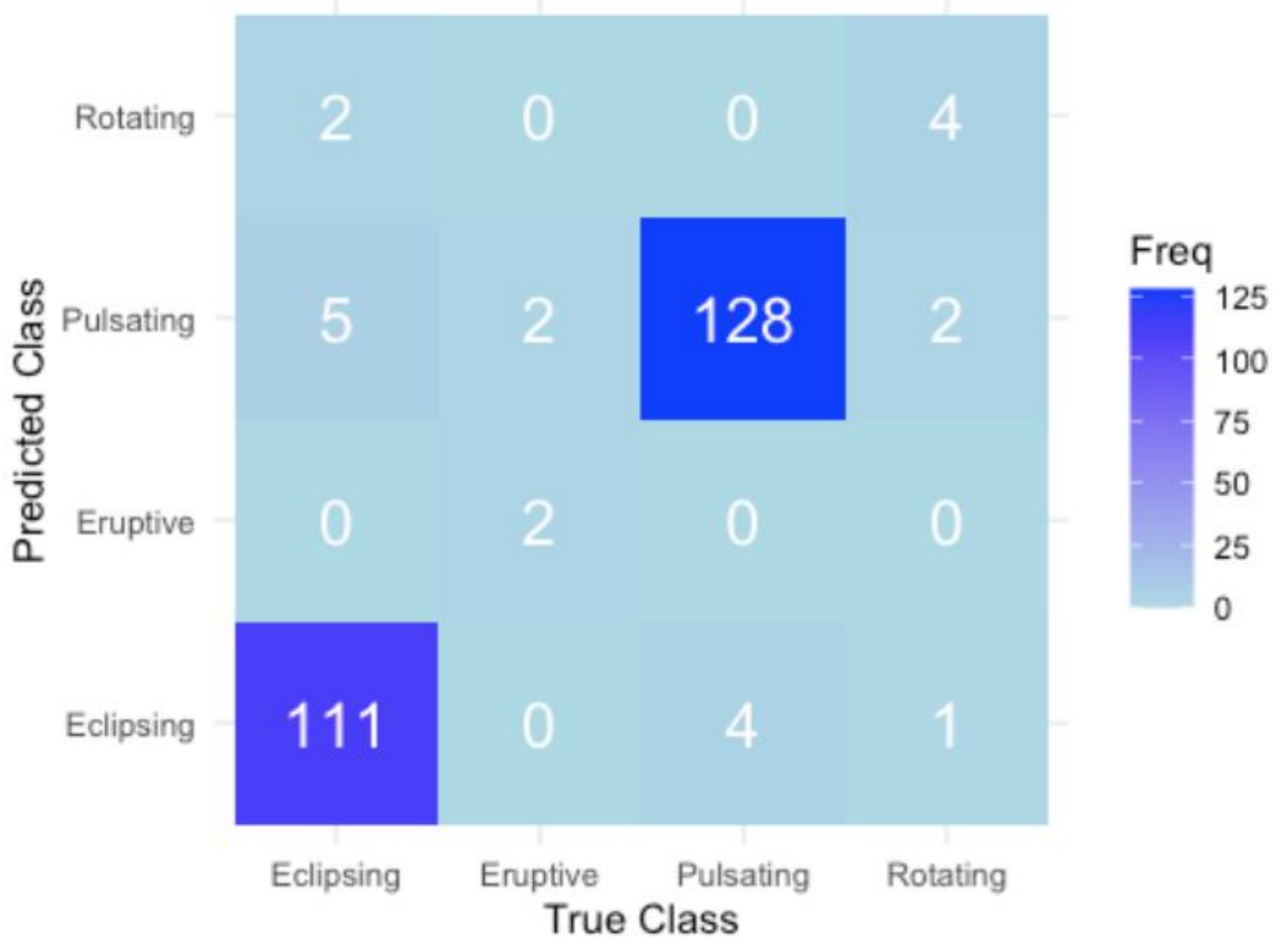


Figure 3: Confusion matrix with the results of the classification performed with the transformer model implementing the BERT architecture, for each of the classes under study.

Conclusions

The implementation of transformer architectures, in particular the BERT architecture adapted to OMC light curves from INTEGRAL, have shown the ability to extract features that allowed their classification in classes with an accuracy of around 94 %, denoting a high effectiveness for future classification of objects studied by this survey. Although preliminary, this result allows us to validate the possibility of implementing such algorithms for the future set of OMC variable sources sources, seeking to improve some conditions such as the balance between the representatives of each variability type. As a next step, our objective is that our classification algorithm is able to distinguish between some of the variability subclasses, such as: Algol, Beta Lyrae, and W UMa eclipsing binaries; Cepheids, Delta Scutis, RR Lyraes, Miras, Long Period Variables, Semi-regular variables between the pulsating stars; Ellipsoidal, BY Draconis, and Alpha2 Canum Venaticorum variables, which are classical rotating stars; and some periodic Orion variables, and T Tauris, which are eruptive stars whose periodic variability is also due to rotation. Additionally, there are other transformer architectures tested such as TST (TiSeries Transformer), Transformer XL, Reformer, or Longformer, which may allow us to find the approach that most efficiently captures the characteristics of the light curves under study.

References

- J. Alfonso-Garzón, A. Domingo, J. M. Mas-Hesse, and A. Giménez 2012. A&A 548.
- Chaini, S., Mahabal, A., Kembhavi, A., & Bianco, F. B. 2024. Astrophysics Source Code Library, 2403.
- Devlin, J. 2018. arXiv preprint arXiv:1810.04805.
- Kenton, M. W. C., Kristina, L., & Devlin, J. 2017. arXiv preprint arXiv:1810.04805.
- Mas-Hesse, J. M., Giménez, A., Culhane, J. L., et al. 2003, A&A, 411, L261
- Morvan, M., Nikolaou, N., Yip, K. H., & Waldmann, I. (2022). arXiv preprint arXiv:2207.02777.
- Ozturk, M. E., Wang, W., Szankin, M., & Shao, L. (2017). Science, 11(5), 746-761.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). arXiv preprint arXiv:1706.03762, 10, S0140525X16001837.

Activity funded by Agencia Estatal de Investigación (Spanish State Research Agency - AEI) MICIU/AEI/10.13039/501100011033 grants PID2019-107061GB-C61 (INTEGRAL) and PID2020-112949GB-I00 (Spanish Virtual Observatory).

MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

AGENCIA ESTATAL DE INVESTIGACIÓN

* E-mail: john.Aguilar@unimilitar.edu.co