Boletim da Sociedade Astronômica Brasileira, **37**, no. 1, 85-86

Machine learning identification of Be stars in photometric surveys

Clara Amorim Navarro[1] & Alex Cavaliéri Carciofi[1]

[1] Instituto de Astronomia, Geofísica e Ciências Atmosféricas da Universidade de São Paulo - IAG/USP
e-mail: clara.navarro@usp.br, carciofi@usp.br

Abstract. Classical Be stars are fast-rotating B-type stars with episodic circumstellar disks, typically identified via spectroscopy—a method incompatible with large-scale detection. This work investigates Be star identification through supervised machine learning applied to ∼ 3,000 OGLE light curves. We compared traditional models using numerical variability features (81–86% accuracy) with Convolutional Neural Networks (CNNs) using light-curve images (88% accuracy). Additionally, an exploratory multiclass CNN for simultaneous identification and orientation inference achieved 71% accuracy. These results demonstrate that machine learning is a viable, promising tool for photometric classification in upcoming surveys like the Vera C. Rubin Observatory (LSST).

Resumo. Estrelas Be clássicas são estrelas do tipo B de alta rotação com discos episódicos, tradicionalmente identificadas por espectroscopia, método incompatível com detecção em larga escala. Este trabalho investiga a identificação de estrelas Be via aprendizado de máquina supervisionado em ∼ 3.000 curvas de luz do OGLE. Comparamos modelos tradicionais baseados em descritores numéricos (acurácia de 81–86%) com Redes Neurais Convolucionais (CNNs) que operam sobre imagens das curvas (88% de acurácia). Uma CNN multiclasse exploratória para identificação e inferência de orientação obteve 71% de acurácia. Os resultados mostram que o aprendizado de máquina é uma via viável e promissora para a classificação fotométrica em grandes levantamentos futuros, como o do Observatório Vera C. Rubin (LSST).

Keywords. Stars: emission line, Be – Techniques: photometric – Methods: data analysis

1. Introduction

Classical Be stars are non-supergiant B-type stars that exhibit, at least once, Balmer emission lines originating from a gaseous circumstellar disk, whose formation and dissipation generate distinctive patterns of photometric variability. While spectroscopy is the most reliable method to identify Be stars, it is observationally expensive and unscalable.

Modern photometric surveys such as the *Optical Gravitational Lensing Experiment* (OGLE, Udalski *et al.* 1992) provide long-term, high-cadence light curves for millions of stars, enabling variability-based studies. In particular, Figueiredo *et al.* (2025) visually classified approximately 3,000 OGLE I-band light curves, identifying 1,751 Be star candidates based on their photometric behavior and color-magnitude evolution.

Motivated by the availability of a labeled sample, this work investigates whether Be stars can be reliably identified using supervised classification methods applied to photometric light curves.

2. Data and Methods

We use OGLE I-band light curves manually analyzed by Figueiredo *et al.* (2025). For each star, this analysis provided a classification (Be or non-Be candidate) and an inferred system orientation (pole-on, edge-on, or unclear), based on disk-related variability and color criteria.

Two distinct approaches were explored for the classification problem. For the first, feature-based approach, we trained supervised classifiers – Random Forest (RF), eXtreme Gradient Boosting (XGBoost), k-Nearest Neighbors (KNN), Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) – using numerical descriptors (features) extracted from the light curves. The feature set includes basic descriptive statistics, such as the median and median absolute deviation (MAD); photometric correlation indices, namely the Stetson J and K indices; temporal information quantified through temporal variogram-based features; frequency-domain descriptors derived from the Lomb-Scargle periodogram; and distribution-shape features based on octiles, including the octile skewness (OS) and the left and right octile weights (LOW/ROW). All features were standardized to zero mean and unit variance prior to training. The second approach was based on convolutional neural networks (CNNs), which were trained using images of the light curves. Two CNN models were implemented: a binary classifier designed to distinguish Be from non-Be stars, and an exploratory multiclass model aimed at simultaneously classifying Be stars and their inferred orientation (pole-on vs. edge-on). The images were generated as black-and-white magnitude vs. time plots, with axes and labels removed. Tick marks were included on the magnitude axis at intervals of 0.1 mag to provide a consistent visual scale to the model, and a common temporal scale was adopted reflecting the observational phases of the OGLE survey.

3. Results and Discussion

3.1. Performance of feature-based classifiers

The performance of the feature-based supervised classifiers is summarized in Tab. 1. Overall accuracies range from 81% to 86%, with RF and XGBoost achieving the best results.

Table 1. Feature-based classifiers performance metrics.

Classifier	Accuracy	Precision	Recall	F1-score
RF	0.86	0.85	0.93	0.89
XGBoost	0.85	0.86	0.92	0.89
KNN	0.84	0.85	0.90	0.87
SVM	0.81	0.84	0.85	0.85
MLP	0.85	0.85	0.92	0.88

The balanced performance between precision and recall throughout the classifiers indicates a robust discrimination between Be and non-Be stars based on photometric variability features alone. This is further illustrated by the confusion matrix for the RF classifier, shown in Fig. 1. The model produced only 17 false negatives in the test set, indicating that the vast majority of Be stars were successfully recovered by the classifier.

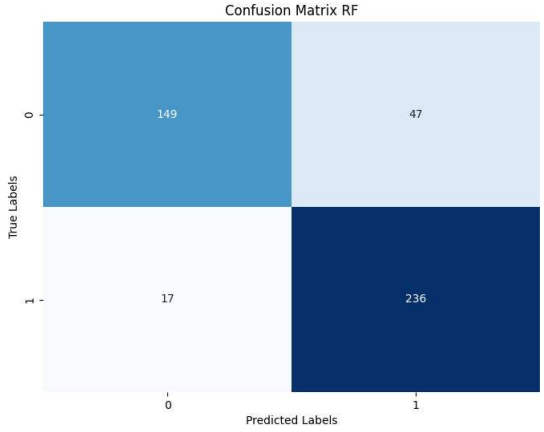

Figure 1. Confusion matrix for the binary classification using the Random Forest (RF) model. Class 0 corresponds to non-Be stars, and Class 1 to Be stars.

To better understand the source of this performance, we analyze the permutation-based feature importance derived from the RF model, shown in Fig. 2. Variogram-based features, as well as the Stetson J index, rank among the most relevant descriptors. This result indicates a key role for temporal information in distinguishing Be stars.

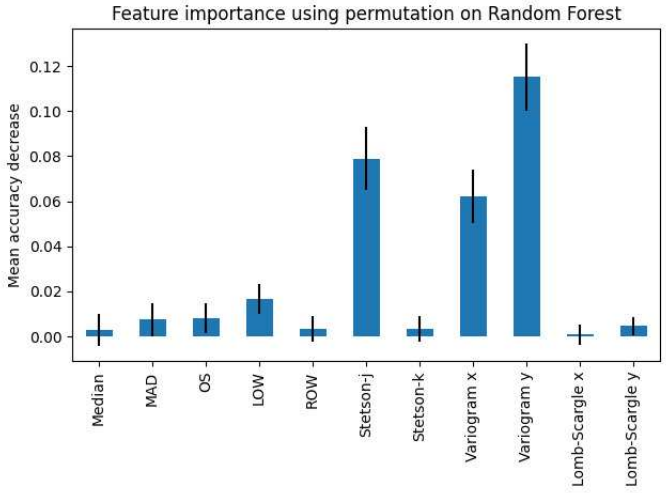

Figure 2. Feature importance for the Random Forest classifier, computed by permuting each feature and measuring the corresponding decrease in model accuracy.

3.2. Performance of convolutional neural networks

The results obtained with convolutional neural networks are reported in Tab. 2. The binary CNN, trained to distinguish Be from non-Be stars, achieves an accuracy of 88%, slightly outperforming the feature-based classifiers. This result demonstrates that the visual morphology of the light curve contains sufficient information for this classification, and that a CNN can effectively extract these morphological signatures without relying on numerical, manually selected features.

A multiclass CNN was also trained to simultaneously classify Be stars and their inferred orientation (pole-on vs. edge-on). The overall accuracy of this model is 71%, with reduced performance primarily driven by two factors: class imbalance on our dataset, and the increased complexity of the classification task.

Table 2. Convolutional neural networks performance metrics.

Classifier	Accuracy	Precision	Recall	F1-score
Binary CNN	0.88	0.87	0.94	0.90
Multiclass CNN	0.71	0.63	0.71	0.63

3.3. Comparison between approaches

Both the feature-based and CNN-based approaches achieve comparable performance. Feature-based models benefit from lower computational cost and straightforward interpretability, whereas CNNs eliminate the need for manual feature engineering by operating directly on light-curve images. This increased flexibility, however, comes at the expense of higher model complexity and training cost. Future applications should take these trade-offs into account when selecting classification strategies, particularly in the context of large-scale photometric surveys.

4. Conclusions

In this work, we show that supervised machine learning methods can successfully identify classical Be stars using photometric data alone. The results indicate that both variability-based descriptors and image-based representations of light curves contain sufficient information for reliable classification, supporting the use of automated approaches for Be star identification.

These findings highlight the feasibility of photometric searches for Be stars in current and upcoming time-domain surveys. The methodology developed in this work provides a foundational framework for future large-scale surveys such as the Vera C. Rubin Observatory LSST, where spectroscopic follow-up for all candidates will be unfeasible.

Acknowledgements. We thank the São Paulo Research Foundation (FAPESP) for funding this project under grant number 2023/1720-0.

References

Figueiredo, A. L., A. C. Carciofi, J. Labadie-Bartz, M. L. Pinho, T. H. de Amorim, P. T. dos Santos, I. Soszynski, A. Udalski 2025, The Astrophysical Journal, 994, 58

Udalski, A., M. Szymanski, J. Kaluzny, M. Kubiak, M. Mateo 1992, Acta Astronomica, 42, 253