Boletim da Sociedade Astronômica Brasileira, **37**, no. 1, 20-21

Use of AI for exoplanet detection and its importance for future discoveries

A. D. Bessa & V. A. Oliveira

[1] Instituto de Física, Universidade de Brasília. e-mail: alicebessadias13@gmail.com

Abstract. Exoplanet detection via direct imaging and high-contrast spectroscopy remains a technically demanding task due to intrinsically low signal-to-noise ratios and the dominance of stellar speckles and instrumental systematics. Traditional techniques based on cross-correlation often struggle to reliably extract weak planetary signatures from these contaminated datasets. This study evaluates the applicability of artificial intelligence, particularly convolutional neural networks (CNNs), to improve the detection of exoplanets in synthetic medium-resolution integral field spectroscopy (IFS) data. Our methodology includes the generation of physically motivated atmospheric models, construction of realistic synthetic IFS cubes, development and training of CNN architectures, and a systematic comparison with classical cross-correlation methods using quantitative performance metrics. By assessing detection efficiency, robustness against false positives, and generalization capability, this work aims to determine the practical advantages of deep-learning approaches for high-contrast spectroscopy. Beyond their technical impact, the results intend to demonstrate that AI-based methods can be effectively implemented within undergraduate research settings, thereby contributing to the advancement of computational astrophysics in Brazil and supporting future discoveries enabled by next-generation instruments.

Resumo. A detecção de exoplanetas por imageamento direto e espectroscopia de alto contraste permanece um desafio técnico devido às baixas relações sinal-ruído e à presença de speckles estelares e artefatos instrumentais. Métodos tradicionais baseados em correlação cruzada frequentemente apresentam limitações para extrair assinaturas planetárias fracas em meio a esses efeitos. Este estudo avalia a aplicabilidade de técnicas de inteligência artificial, em especial redes neurais convolucionais (CNNs), para aprimorar a detecção de exoplanetas em dados sintéticos de espectroscopia de campo integrado (IFS) de resolução média. A metodologia envolve a geração de modelos atmosféricos fisicamente consistentes, construção de cubos IFS sintéticos realistas, desenvolvimento e treinamento das arquiteturas de CNN e uma comparação sistemática com técnicas clássicas de correlação cruzada por meio de métricas quantitativas. Ao analisar eficiência de detecção, redução de falsos positivos e capacidade de generalização, busca-se demonstrar as vantagens práticas de abordagens baseadas em deep learning. Além de seu impacto técnico, o trabalho reforça que metodologias de IA podem ser implementadas em pesquisas de graduação, contribuindo para o avanço da astrofísica computacional no Brasil e para futuras descobertas com instrumentos de próxima geração.

Keywords. Planets and satellites: detections, data analysis, Gravitational lensing: micro

1. Introduction

The detection of exoplanets is one of the central challenges of modern astrophysics, essential for understanding the formation of planetary systems and for investigating the possibility of life beyond the Solar System. Traditional methods such as transit, radial velocity, gravitational microlensing, and high-contrast spectroscopy present limitations due to instrumental noise, low signal-to-noise ratio, and stellar contamination. In this context, Artificial Intelligence (AI) techniques emerge as promising tools capable of enhancing the identification of planetary signals, reducing false positives, and handling large volumes of observational data.

2. Methodology

The methodology of this work is structured into five main stages: data acquisition, preprocessing, machine learning application, method integration, and validation. Each stage is detailed below.

2.1. Data Acquisition

The database for this study comprises two primary types of data:

- Photometric Time Series: Real light curves from the Kepler and TESS missions, which form the basis for transit detection algorithms.
- Synthetic Spectroscopic Data: Realistically simulated integral field spectroscopy (IFS) cubes, generated to train and test models for high-contrast, direct imaging applications.

These datasets encompass three distinct classes: positives (containing a planetary transit signal), negatives (no transit), and false positives (signals mimicked by instrumental or astrophysical noise).

2.2. Preprocessing

To ensure data quality and compatibility with machine learning algorithms, all time series and spectral data underwent a standardized preprocessing pipeline:

- Noise Removal: Application of filtering techniques (e.g., Gaussian smoothing, Savitzky-Golay filters) to reduce high-frequency instrumental and photon noise.
- Normalization: Flux values were normalized to a common baseline, facilitating comparison and feature extraction.
- Gap Treatment: Missing data points or gaps in the time series were handled via interpolation or masking to prevent artifacts in the analysis.
- Windowing: Data were segmented into standardized windows, focusing analysis on regions of interest (e.g., around expected transit events).

2.3. Machine Learning Approaches

Two complementary strategies were implemented to classify the preprocessed data:

2.3.1. Global Approach

This method considers the entire light curve or spectrum as a single feature vector. Nine classical and ensemble models were trained and evaluated

2.3.2. Local Approach

This technique focuses the analysis solely on the localized region where a transit signal is suspected, increasing sensitivity to subtle flux variations. Three high-performing ensemble models were adapted for this task

- Logistic Regression: A linear probabilistic classifier providing a baseline for interpretability.
- K-Nearest Neighbors (KNN): A non-parametric, instance-based learning algorithm.
- Naive Bayes: A probabilistic classifier based on Bayes' theorem with feature independence assumptions.
- Decision Tree: A flowchart-like model that learns simple decision rules from the data features.
- Support Vector Machine (SVM): A powerful classifier that finds the optimal separating hyperplane in high-dimensional space.
- Random Forest: An ensemble of decision trees that reduces overfitting through bagging and feature randomness.Leveraged for its robustness and ability to capture complex patterns in the local transit morphology
- XGBoost: An optimized gradient boosting framework known for its speed and performance. Utilized for its precision in modeling sequential dependencies within the transit window
- LightGBM: A gradient boosting framework designed for efficiency and scalability with large datasets. Employed for its computational efficiency when processing high-resolution local features.
- CatBoost: A gradient boosting algorithm effective at handling categorical features with minimal preprocessing.

Additionally, the implementation of Convolutional Neural Networks (CNNs) for analyzing two-dimensional spectral-spatial data (IFS cubes) is planned to extend the methodology to high-contrast spectroscopic applications.

2.4. Validation

The performance and reliability of all models are rigorously assessed through:

- Benchmark Comparison: Systematic comparison of results against established benchmarks and published findings in the literature.
- Quantitative Performance Metrics: Evaluation using standard metrics such as accuracy, precision, recall, F1-score, and the area under the Receiver Operating Characteristic (ROC-AUC) curve. The reduction of false positives is a key metric of interest.

3. Preliminary Results

The algorithms demonstrated the ability to identify light curves with real transits even in scenarios of strong noise, with consistent performance across the classes. The presence of false positives was reduced especially in approaches based on *ensemble methods*. These results reinforce the potential of using AI for exoplanet detection in photometric time series.

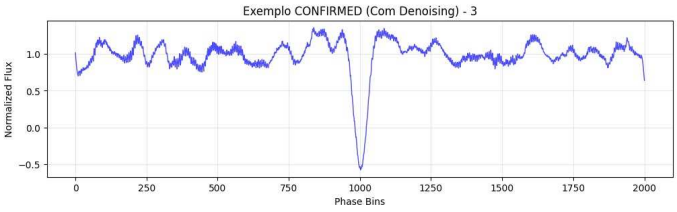

Figure 1. Light curve classified as *CONFIRMED* after the *denoising* process.

4. Future Work

The next steps include:

- Developing a hybrid database for coupled methods: Initially, we will focus on constructing a comprehensive dataset to enable future testing of the coupling between planetary transit and gravitational microlensing detection techniques.
- Implementing CNNs for the analysis of integral field spectroscopy.
- Expanding the database with atmospherically realistic simulations.
- Comparing performance with classical cross-correlation techniques.

These actions aim to consolidate the use of AI in Brazilian astronomy, increasing precision in planetary detection and preparing the ground for observations with next-generation instruments.

5. Final Considerations

The preliminary results confirm that AI algorithms are effective in classifying light curves containing planetary signals, even under noisy conditions. This work demonstrates that advanced techniques can be developed at the undergraduate level, promoting integration between astronomy and data science and contributing to future discoveries in the field of exoplanets.

6. References

Garvin, E. O., Bonse, M. J., Hayoz, J., et al. 2025, Machine Learning for Exoplanet Detection in High-Contrast Spectroscopy (em preparação)
Macedo, B. H. D. & Zalewski, W. 2023, Rev. Bras. Iniciação Científica, e024021
Rithivraj, G. & Kumari, A. 2023, arXiv e-prints, arXiv:2305.05956