Overcoming machine learning training data imbalance by simulating exoplanet transits

Nika Gorchakova,[1] Oisín Creaner [1]

[1]*Dublin City University, Dublin, Ireland;* nika.gorchakova2@mail.dcu.ie

Abstract. We propose to use simulations of exoplanet transits to improve training outcomes for Machine Learning models. Machine learning has huge potential in exoplanet detection but faces challenges due to data imbalance and lack of ground truth in observational data. Most stars do not show transits, leading to datasets being skewed towards non-transit light curves, which can result in over-fitting and poor recall. Furthermore, the absence of ground truth complicates understanding the effects of noise and errors on detection outcomes. To address these issues, we simulate exoplanet transits using key astrophysical parameters and diverse noise profiles to create balanced training datasets. This simulation-based approach will improve machine learning models, enhancing their outcomes in detecting exoplanets in real-world data.

1. Introduction

Exoplanet transits are observed as a periodic dimming of the host star when an orbiting planet passes in front of it (Henry et al. 1999). Observations capture a series of images or frames over time. These frames are processed to extract the brightness (flux) of the target star. This data is then used to create a light curve, a graph showing variations in the star's brightness. Finally, the light curve is fitted with models to determine the planet's characteristics. This process is represented in Figure 1.

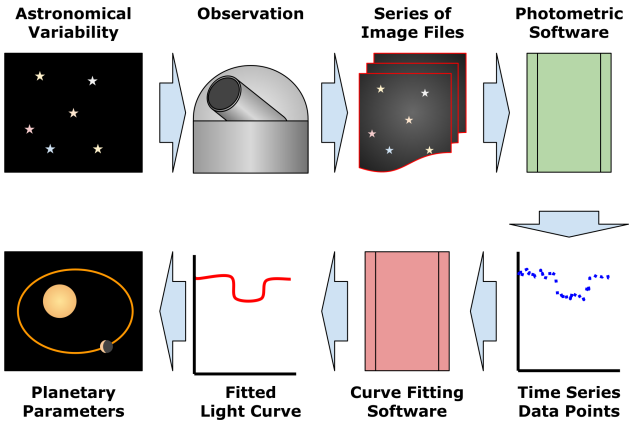

Figure 1. Schematic of an idealised exoplanet transit observation pipeline.

Machine learning (ML) offers transformative potential for exoplanet detection by enabling the rapid and automated analysis of vast astronomical datasets (Sen et al. 2022). ML models excel at identifying patterns and extracting meaningful signals from noisy, complex data, making them particularly effective for detecting the subtle signatures of exoplanets(Cuéllar et al. 2022). Existing work (Malik et al. 2021; Pimentel et al. 2024) focuses on extracting features from light curves. This project, ExoBytes, goes one step further by applying computer vision techniques to directly extract transits from images.

However, the performance of ML models is highly dependent on the quality and balance of their training datasets. Unfortunately, transits are short, infrequent events, meaning astronomical archives represent highly imbalanced datasets. To address these challenges, we propose a simulation-based approach to generate diverse, balanced datasets. By simulating observations of exoplanet transits with realistic noise profiles, we aim to improve the robustness of ML models for real-world exoplanet detection.

2. Training Problem

Identifying a transit in astronomical images is an example of a classification task: separating the observations into categories of "transit" and "not a transit". Supervised Machine Learning (SML) techniques are most commonly used for classification. SML relies on training data: sets of input data which has already been categorised (Mohri 2018). However, issues arise if the training data is *imbalanced* (one category outnumbers the other — Megahed et al. (2021)) or if the classification labels are *imperfect* (the ground truth of the underlying data is uncertain — Cannings et al. (2020)). Real astronomical images suffer from both sources of error.

2.1. Data Imbalance

In any given set of astronomical images, exoplanet transits are vastly outnumbered by non-transiting stars.

Such imbalance can cause a model to become biased toward the majority class Megahed et al. (2021) — in this case, we would expect poor performance in identifying transits. Existing research often addresses this issue by oversampling the minority class with techniques such as Synthetic Minority Oversampling Technique (SMOTE), which creates synthetic samples by interpolating between existing minority class samples (Chawla et al. 2002). However, this approach can amplify noise or outliers, potentially leading to over-fitting Megahed et al. (2021).

2.2. Ground Truth

Exoplanet observations, even from spacecraft, are inherently subject to substantial error and uncertainty (Thompson et al. 2018). This leads to a lack of definitive ground truth in exoplanet data. That further complicates model training and evaluation (Cannings et al. 2020). This uncertainty imposes challenges when assessing model performance and when isolating the impact of noise and errors on detection outcomes. As shown in section 3, we aim to eliminate this uncertainty using artificial signals to develop of reliable detection algorithms. However, we also underscore the need for meticulous validation against natural exoplanet signals for robustness in practical applications.

3. Solution

ExoBytes proposes to simulate time-series images with balanced samples of transits and non-transiting stars. These will be used in training machine learning systems.

3.1. Simulating images

A central aspect of this approach is the simulation of time-series images rather than light curves. Simulated images closely replicate raw observational data, capturing both spatial and contextual features that are essential for robust ML training. While there are many solutions for simulating individual images, efficiently generating large volumes of images remains challenging due to the high computational demands, the need for precise astrophysical modelling (Peterson 2014), and the inclusion of realistic noise and variability to accurately reflect actual observations. Simulating images instead of light curves offers several advantages:

More Realistic Data Representation: Simulated images reflect real-world observational conditions, including noise and artifacts, making ML models more robust for practical applications (Cuéllar et al. 2022).

Enhanced Noise Handling: Training models on visual datasets helps them better differentiate between genuine transit signals and false positives caused by noise or artifacts (Sarkar et al. 2021).

Improved Feature Extraction: Images preserve spatial features, which are lost in one-dimensional light curves (Carrasco-Davis et al. 2019). This enhances the model's ability to identify and process subtle or irregular patterns.

Controlled environment: Simulations provide a controlled environment with known ground truth, enabling precise calibration of machine learning models and offering a reliable platform to test and validate their performance before deployment in observational settings (Sarkar et al. 2021).

Balanced datasets: With simulation we can generate dataset with equal numbers of transit and non-transit examples. This enables ML models to generalize better across rare transit signals and common non-transit cases(Braga et al. 2022).

3.2. Simulation process

This section outlines the key steps required to produce simulated images. In some ways, this is the inverse of the process to observe exoplanet transits illustrated in Figure 1.

Defining Astronomical Parameters: The characteristics of the exoplanet and host star such as radii, period, inclination etc. are specified.

Generating the Light Curve: Utilizing the defined parameters, a ideal light curve is produced, representing the star's brightness variations over time due to the transit.

Star field generation: A detailed model of the star field, including the density and magnitude distribution of stars is created.

Atmospheric Effects: The point spread function (PSF) is configured to account for effects like atmospheric seeing and star blending. Additionally, atmospheric noise effects are distributed across the field.

Detector Effects: Detector effects from both telescope (e.g. diffraction) and camera (e.g. read noise) are included.

Image Generation: FITS images are generated for key transit phases, including ingress, egress, and out-of-transit baselines.

Validation: Light curves are extracted and compared them to the original models.

4. Conclusion

Simulating exoplanet transits offers a promising solution to the challenges of data imbalance and the absence of ground truth in machine learning (ML) for exoplanet detection. By generating large and diverse datasets, simulations address the limitations of observational data, enabling ML models to learn from a balanced representation of transit and non-transit cases. The controlled environment provided by simulations facilitates precise calibration and systematic evaluation of model performance, helping to isolate and understand the impact of various noise sources.

Simulated datasets train ML models to handle a wide range of noise profiles, enhancing their robustness to real-world variability. Furthermore, they prepare models for novel or rare cases that are underrepresented in observational datasets. The ability to rigorously test and validate models using simulated data ensures that the models perform reliably under diverse and challenging conditions.

4.1. Next Steps

Once established, the simulated images will be integrated into ML training datasets, enriching them with diverse and balanced examples. The resulting models will be tested on real-life datasets to evaluate their accuracy, robustness, and ability to handle various noise profiles. To evaluate the effectiveness of the proposed approach, a comparative analysis will be conducted to assess the performance of models trained on datasets enriched with simulated data versus those trained using selective sampling techniques.

Acknowledgments. ExoBytes (NG and OC) are funded by Science Foundation Ireland through the SFI Centre for Research Training in Machine Learning (Grant No. 18/CRT/6183) Call 2023. Star Guide (OC) is funded by the National Open Research Forum (NORF) Open Research Fund 2023, Strand II: Open Research Stimulus. This funding is provided through the Higher Education Authority (HEA).

References

Braga, F. C., Roman, N. T., & Falceta-Gonçalves, D. 2022, in Intelligent systems, edited by J. C. Xavier-Junior, & R. A. Rios (Cham: Springer International Publishing), 107
Cannings, T. I., Fan, Y., & Samworth, R. J. 2020, Biometrika, 107, 311
Carrasco-Davis, R., et al. 2019, PASP, 131, 108006
Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. 2002, JAIR, 16, 321
Cuéllar, S., Granados, P., Fabregas, E., Curé, M., Vargas, H., Dormido-Canto, S., & Farias, G. 2022, PLOS ONE, 17, e0268199
Henry, G. W., Marcy, G. W., Butler, R. P., & Vogt, S. S. 1999, ApJ, 529, L41
Malik, A., Moster, B. P., & Obermeier, C. 2021, MNRAS. Publisher: Oxford University Press (OUP)
Megahed, F. M., Chen, Y.-J., Megahed, A., Ong, Y., Altman, N., & Krzywinski, M. 2021, Nat. Methods, 18, 1270
Mohri, M. 2018, Foundations of machine learning
Peterson, J. R. 2014, JINST, 9, C04010
Pimentel, J., Amorim, J., & Rudzicz, F. 2024, Int. J. Data Sci. Anal.
Sarkar, S., Pascale, E., Papageorgiou, A., Johnson, L. J., & Waldmann, I. 2021, Exp. Astron., 51, 287
Sen, S., Agarwal, S., Chakraborty, P., & Singh, K. P. 2022, Exp. Astron., 53, 1
Thompson, S. E., et al. 2018, ApJS, 235, 38