CLARA: A Modular Framework for Unsupervised Transit Detection Using TESS Light Curves

M.Dasgupta,[1]★
[1]*Indian Institute of Technology, Madras*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

We present CLARA, a novel framework for unsupervised exoplanet detection in TESS light curves through systematic optimization of Unsupervised Random Forest (URF) models. Our research addresses two fundamental questions:

(a) How synthetic training set design affects URF performance and generalization across independent TESS SPOC 2-minute cadence sectors, and

(b) Whether URF anomaly scores correlate with genuine astrophysical phenomena for filtering transiting objects of interest.

In Part I, we investigate synthetic dataset construction impact on URF performance across diverse TESS light curve morphologies. We introduce a parameterized scoring framework controlled by a single variable (0 1) that balances precision and recall metrics, enabling quantitative generalization assessment across stellar populations and observational conditions. Carefully designed synthetic sets significantly enhance detection efficiency while maintaining robust cross-sector generalization.

In Part II, we validate URF astrophysical significance through correlation analysis between model scores and known stellar/planetary properties. Using improved Normalized Weighted Root Sum of Squares (NWRSS) scoring, we demonstrate strong correlations between high anomaly scores and genuine astrophysical phenomena via t-SNE clustering. We establish morphological classification pipelines that filter light curves based on SIMBAD catalog similarity for select TOIs, providing quantitative validation that URF models detect astrophysical signals rather than instrumental artifacts.

Our methodology demonstrates that properly optimized unsupervised approaches significantly exceed supervised detection performance, opening new avenues for large-scale exoplanet discovery in archival survey data. CLARA's modular design enables scalable processing of entire TESS sectors while maintaining interpretable connections between computational anomaly detection and physical astrophysical processes.

Key words: exoplanets – time series analysis – unsupervised learning – TESS – anomaly detection – synthetic datasets – morphological classification

1 INTRODUCTION

The discovery and characterization of astrophysical phenomena from large-scale time-series surveys represents one of the most challenging problems in modern astronomy. With missions like the Transiting Exoplanet Survey Satellite (TESS) generating millions of light curves, traditional supervised approaches for anomaly detection face fundamental limitations due to the rarity and diversity of interesting astrophysical events . The need for scalable, unsupervised methods that can systematically discover and interpret anomalies—such as exoplanet transits, eclipsing binaries, and novel variable stars—has become increasingly urgent as survey data volumes continue to grow exponentially.

Recent advances in unsupervised machine learning for astronomical time-series analysis have shown promising results. Nun et al. (2015) demonstrated the effectiveness of Random Forest-based approaches for stellar variability classification, while Malanchev et al. (2021) explored deep learning methods for supernova discovery in large surveys. Active learning approaches have also emerged as powerful tools for personalized anomaly detection, with Lochner & Bassett (2016) introducing frameworks that combine human intuition with machine learning efficiency. The application of Unsupervised Random Forests (URFs) to exoplanet detection was pioneered by Crake & Martínez-Galarza (2023), who showed that synthetic training sets could guide anomaly detection behavior in TESS light curves. However, these approaches have lacked systematic frameworks for controlling the discovery process and interpreting the physical nature of detected anomalies.

The challenge of anomaly detection in astrophysical time-series data is compounded by several factors: the high dimensionality of light curve features, the presence of various noise sources and systematic effects, and the need to distinguish between different classes of astrophysical phenomena Ivezić et al. (2019). Traditional clustering and dimensionality reduction techniques often fail to provide physically meaningful groupings of astronomical objects Ivezić et al. (2019), while purely data-driven approaches may discover statisti-

★ E-mail: senguptasumitra48@gmail.com

cally significant but astrophysically uninteresting patterns. Moreover, the analysis of time series photometry requires Lomb-Scargle periodogram techniques VanderPlas (2018). to properly characterize the temporal variability of astronomical sources that have been unevenly sampled.

We present CLARA (Controllable Learning for Anomaly Recognition in Astrophysics), a comprehensive pipeline that addresses these limitations through systematic synthetic set design and feature-to-metric mapping Crake & Martínez-Galarza (building upon 2023). Our approach introduces novel methodologies for steering URF behavior through controlled variation of synthetic training parameters, enabling the targeted discovery of specific astrophysical phenomena. Furthermore, we develop morphological classification techniques using cosine similarity matching with known objects, and integrate astrometric data from Gaia DR3 Gaia Collaboration et al. (2023) to provide physical interpretation of discovered anomalies. This work aims to transform unsupervised anomaly detection from a purely statistical exercise into a physically guided discovery tool for astronomical surveys.

2 EXPLORING HOW THE DESIGN OF SYNTHETIC SETS AFFECTS UNSUPERVISED ANOMALY DETECTION

Unsupervised random forest models for anomaly detection were first used in astronomical datasets in Baron & Poznanski (2017) where isolation forests were applied to SDSS galaxy data to identify the most unusual and potentially rare objects in the survey. As demonstrated ahead, the controllability of unsupervised anomaly detection hinges on our ability to systematically design synthetic training sets that guide model behavior toward discovering specific types of astrophysical phenomena. By varying key parameters such as the number of synthetic curves in the fake data set, duration of curves/transits, frequency, and noise characteristics in our synthetic light curves, we can **steer Unsupervised Random Forests** to prioritize different anomaly signatures and effectively "tune" the discovery process for targeted scientific objectives.

2.1 Feature of a Curve

The characteristic of a single light curve for input to our URF (Unsupervised Random Forest) model experiments follows established methods from our foundational paper Crake & Martínez-Galarza (2023) where the first n=3000 flux points are stacked on top of nls = 1000 Lomb-Scargle periodogram, VanderPlas (2018) frequency points calculated using the LombScargle function of the astropy.timeseries module. The feature is thus an array with 4000 values. This function calculates the power of a time period where the power describes the extent of the periodicity of the selected time value, given by:

$$P_{LS}(\omega) = \frac{1}{2}\left[\frac{\left(\sum_j (x_j - \bar{x})\cos(\omega(t_j - \tau))\right)^2}{\sum_j \cos^2(\omega(t_j - \tau))} + \frac{\left(\sum_j (x_j - \bar{x})\sin(\omega(t_j - \tau))\right)^2}{\sum_j \sin^2(\omega(t_j - \tau))}\right] \quad (1)$$

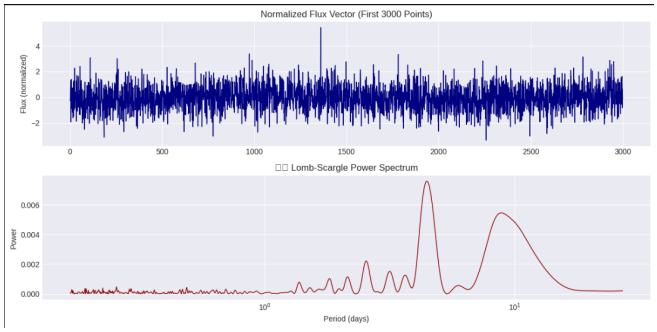

Figure 1. The first 3000 flux points of a light curve (top plot) and the Lomb-Scargle periodogram of the same curve (bottom plot)

where:

$P_{LS}(\omega)$ = Lomb-Scargle power at angular frequency ω
x_j = observed flux values at times t_j
$\bar{x}$ = mean of the observed flux values
τ = time offset defined by:

$$\tan(2\omega\tau) = \frac{\sum_j \sin(2\omega t_j)}{\sum_j \cos(2\omega t_j)}$$

2.2 URF Hyperparamter Randomized Search via Real + Synthetic Curve Stacking

For the real set (real features class), we took features of 1500 real curves from sector 2. A `RandomizedSearchCV` is performed over a predefined hyperparameter space for the `RandomForestClassifier`. The search is:

- Randomized (`n_iter=30`)
- Parallel (`n_jobs=-1`)
- Reproducible (`random_state=42`)

We use the following hyper-parameter search space for the randomized search:

`n_estimators`: 10 evenly spaced values between 50 and 200
`max_features`: [sqrt, log2]
`max_depth`: [100, 300, 500, 700, 900, 1000, None]
`min_samples_split`: [2, 4, 7, 10]
`min_samples_leaf`: [1, 2]
`bootstrap`: [True, False]
`warm_start`: [True, False]

For our initial tests, we took the following synthetic set (fake features class) design types for the URF.

a. **URF1**. We generated the synthetic set by uniformly sampling feature values within the range of the real dataset. This serves as a baseline method.
X_fake = np.random.uniform(np.min(X_real), np.max(X_real), size=X_real.shape)

b. **URF2**. We used fixed hyper params mentioned at Crake & Martínez-Galarza (building upon 2023). The following hyper params were used:

- `n_estimators = 100`

- `max_features = "sqrt"`
- `max_depth = 700`
- `min_samples_split = 4`
- `min_samples_leaf = 2`
- `bootstrap = False`
- `warm_start = True`

c. **URF3**. We used a TOI based synthetic class where we obtained feature vectors from Tess Objects of Interest (TOIs) from Sector 1 of SPOC data, and set those features as the synthetic set for hyper-parameter search.

d. **URF4**. The synthetic set of features were computed from a set of synthetic light curves generated using transit models (box-shaped model and Mandel-Agol) injected into realistic noise baselines. The computational implementation of this was performed using the batman pypi package.

The four URF variants were designed to systematically explore different approaches to synthetic set construction for transit discovery. URF1 serves as a baseline control, using uniform random sampling within the feature space bounds to establish minimum performance expectations. URF2 tests the generalizability of the hyperparameters established in Crake & Martínez-Galarza (2023), evaluating whether their optimization for stellar variability classification translates effectively to transit detection in TESS data. URF3 explores a realistic "ground truth" approach by using feature vectors from confirmed TOIs as the synthetic class, providing insight into the performance ceiling when the synthetic set closely matches actual transit signatures. Finally, URF4 introduces our controllable framework, where synthetic light curves are generated using physically motivated transit models with systematically variable parameters (ratio of actual curves - 1500 - to synthetic curves - 300), duration, cadence, and noise characteristics), enabling targeted optimization for specific discovery objectives and forming the foundation for our systematic parameter exploration.

2.3 Anomaly Scoring

Following Baron & Poznanski (2017) and Crake & Martínez-Galarza (2023), our URF implementation uses terminal node population as the anomaly scoring heuristic. Each tree in the forest is trained to distinguish real light curves from synthetic ones, with anomaly scores computed by tracking the population composition of terminal nodes. For each real light curve, we calculate a similarity score S as the average fraction of real data points sharing the same terminal node across all trees in the forest. The final anomaly score is defined as 1S, where a score of 1 indicates maximum anomaly (the object consistently lands alone in terminal nodes) and 0 indicates minimal anomaly (the object is always grouped with the majority of real data). This population-based approach provides a natural measure of how isolated each light curve appears relative to the broader dataset distribution.

2.4 Initial Results

Table 1 summarizes the performance of our four URF variants on TOI detection across the test dataset. URF-1, URF-2, and URF-4 demonstrate similar scoring behavior, assigning zero scores to most light curves and treating only non-zero scores as anomalies. URF-1 and URF-2 achieve baseline recalls of 12.6% and 25.7% respectively with ~1% precision, while URF-4 shows improved performance at

Table 1. Performance comparison of URF variants on TOI detection. TP = True Positives, FP = False Positives, TN = True Negatives, with precision and recall calculated for the 175 total TOIs in the test set for URFs 1,2 and 4 and 195 total TOIs in the test set for URF 3

URF Variant	TP	FP	TP+FP	TN	Precision (%)	Recall (%)
URF-1	22	2159	2181	0	1.01	12.57
URF-2	45	3990	4035	0	1.12	25.71
URF-3	14	300	314	15499	4.46	7.18
URF-4	64	4982	5046	0	1.27	36.57

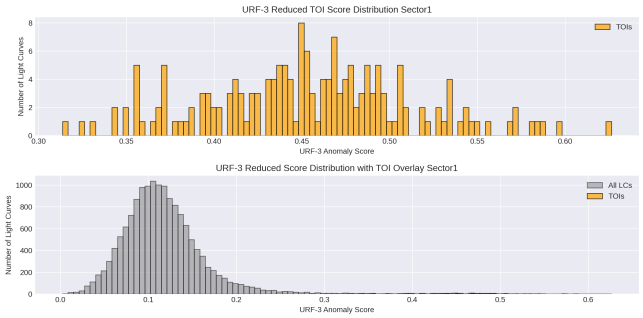

Figure 2. URF3 TOI (top) and All Curves (bottom) Anomaly Score Distribution for Sector 1 Test Set

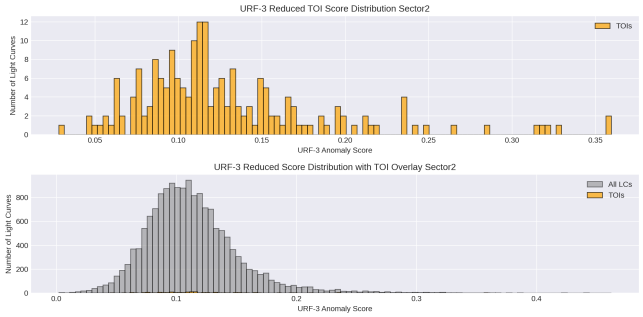

Figure 3. URF3 TOI (top) and All Curves (bottom) Anomaly Score Distribution for Sector 2 Test Set

36.6% recall and 1.3% precision. URF-3 exhibits fundamentally different behavior, assigning non-zero scores to nearly all light curves, requiring threshold-based evaluation (scores > 0.22) to define meaningful anomaly regions. When evaluated under these conditions, URF-3 achieves 7.2% recall with 4.5% precision, though this performance is best assessed on independent sectors to avoid overfitting bias from using TOI features as synthetic training data. The distinct scoring profiles highlight URF-4's advantage in providing both strong recall performance and systematic controllability without the threshold sensitivity issues observed in URF-3

Note on URF-3 Cross-Sector Validation: URF-3's performance exhibits significant sector-dependent variation due to its training methodology. Since URF-3 uses feature vectors extracted from confirmed TOIs in Sector 1 as its synthetic anomaly class, it demonstrates clear overfitting when evaluated on the same sector. This is evident in the score distribution plots (Figures 2 and 3), where

URF-3's anomaly scores for TOIs in Sector 1 are systematically higher (peaking around 0.45-0.47) compared to Sector 2 (peaking around 0.10-0.12). The reduced score separation in Sector 2 reflects URF-3's diminished ability to generalize beyond the specific feature characteristics of its training TOIs. This sector dependency underscores the importance of independent validation sets and highlights a fundamental limitation of using confirmed target features directly as synthetic training data, reinforcing URF-4's advantage in using parametric synthetic generation for improved generalizability.

2.5 Figures and tables

Figures and tables should be placed at logical positions in the text. Don't worry about the exact layout, which will be handled by the publishers.

Figures are referred to as e.g. Fig. **??**, and tables as e.g. Table **??**.

3 CONCLUSIONS

The last numbered section should briefly summarise what has been done, and describe the final conclusions which the authors draw from their work.

ACKNOWLEDGEMENTS

The Acknowledgements section is not numbered. Here you can thank helpful colleagues, acknowledge funding agencies, telescopes and facilities used etc. Try to keep it short.

DATA AVAILABILITY

The inclusion of a Data Availability Statement is a requirement for articles published in MNRAS. Data Availability Statements provide a standardised format for readers to understand the availability of data underlying the research results described in the article. The statement may refer to original data generated in the course of the study or to third-party data analysed in the article. The statement should describe and provide means of access, where possible, by linking to the data or providing the required accession numbers for the relevant databases or DOIs.

REFERENCES

Baron D., Poznanski D., 2017, MNRAS, 465, 4530
Crake D. A., Martínez-Galarza J. R., 2023, MNRAS, 520, 1234
Gaia Collaboration Vallenari A., Brown A. G. A., et al., 2023, A&A, 674, A1
Ivezić Ž., Connolly A. J., VanderPlas J. T., Gray A., 2019, Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data, 2nd edn. Princeton University Press, Princeton, NJ
Lochner M., Bassett B. A., 2016, ApJS, 225, 31
Malanchev K. L., Pruzhinskaya M. V., Korolev V. S., et al., 2021, MNRAS, 502, 5147
Nun I., Pichara K., Protopapas P., Patel P., 2015, ApJ, 793, 23
VanderPlas J. T., 2018, ApJS, 236, 16

APPENDIX A: SOME EXTRA MATERIAL

If you want to present additional material which would interrupt the flow of the main paper, it can be placed in an Appendix which appears after the list of references.

This paper has been typeset from a TEX/LATEX file prepared by the author.