UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Finding radio transients with anomaly detection and active learning based on volunteer classifications

Andersson, Alex; Lintott, Chris; Fender, Rob; Lochner, Michelle; Woudt, Patrick; van van Eijnden, Jakob; van der Horst, Alexander; Horesh, Assaf; Saikia, Payaswini; Sivakoff, Gregory R.; Tremou, Lilia; Vaccari, Mattia

DOI
10.1093/mnras/staf336

Publication date
2025

Document Version
Final published version

Published in
Monthly Notices of the Royal Astronomical Society

License
CC BY

Link to publication

Citation for published version (APA):
Andersson, A., Lintott, C., Fender, R., Lochner, M., Woudt, P., van van Eijnden, J., van der Horst, A., Horesh, A., Saikia, P., Sivakoff, G. R., Tremou, L., & Vaccari, M. (2025). Finding radio transients with anomaly detection and active learning based on volunteer classifications. *Monthly Notices of the Royal Astronomical Society*, *538*(3), 1397-1414. https://doi.org/10.1093/mnras/staf336

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:09 Sept 2026

Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **538,** 1397–1414 (2025) https://doi.org/10.1093/mnras/staf336
Advance Access publication 2025 February 25

Finding radio transients with anomaly detection and active learning based on volunteer classifications

Alex Andersson [1★] Chris Lintott [1] Rob Fender [1,2] Michelle Lochner [3,4] Patrick Woudt [2] Jakob van den Eijnden [5,6] Alexander van der Horst [7] Assaf Horesh [8] Payaswini Saikia [9] Gregory R. Sivakoff [10] Lilia Tremou [11] and Mattia Vaccari [12,13,14]

[1]*Astrophysics, Department of Physics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH, UK*
[2]*Department of Astronomy, University of Cape Town, Private Bag X3, Rondebosch 7701, South Africa*
[3]*Department of Physics and Astronomy, University of the Western Cape, Bellville, Cape Town 7535, South Africa*
[4]*South African Radio Astronomy Observatory, 2 Fir Street, Black River Park, Observatory, Cape Town 7925, South Africa*
[5]*Department of Physics, University of Warwick, Coventry CV4 7AL, UK*
[6]*Anton Pannekoek Institute for Astronomy, Universiteit van Amsterdam, Science Park 904, XH 1098, Amsterdam, The Netherlands*
[7]*Department of Physics, The George Washington University, 725 21st Street NW, Washington, DC 20052, USA*
[8]*Racah Institute of Physics, The Hebrew University of Jerusalem, Jerusalem 91904, Israel*
[9]*Center for Astro, Particle and Planetary Physics, New York University Abu Dhabi, PO Box 129188 Abu Dhabi, UAE*
[10]*Department of Physics, University of Alberta, CCIS 4-181, Edmonton, AB T6G 2E1, Canada*
[11]*National Radio Astronomy Observatory, Socorro, NM 87801, USA*
[12]*Inter-University Institute for Data Intensive Astronomy, Department of Astronomy, University of Cape Town, 7701 Rondebosch, Cape Town, South Africa*
[13]*Inter-University Institute for Data Intensive Astronomy, Department of Physics and Astronomy, University of the Western Cape, Bellville 7535, Cape Town, South Africa*
[14]*INAF – Istituto di Radioastronomia, via Gobetti 101, I-40129 Bologna, Italy*

Accepted 2025 February 21. Received 2025 February 21; in original form 2024 September 30

ABSTRACT

In this work, we explore the applicability of unsupervised machine learning algorithms to finding radio transients. Facilities such as the Square Kilometre Array (SKA) will provide huge volumes of data in which to detect rare transients; the challenge for astronomers is how to find them. We demonstrate the effectiveness of anomaly detection algorithms using 1.3 GHz light curves from the SKA precursor MeerKAT. We make use of three sets of descriptive parameters ('feature sets') as applied to two anomaly detection techniques in the ASTRONOMALY package and analyse our performance by comparison with citizen science labels on the same data set. Using transients found by volunteers as our ground truth, we demonstrate that anomaly detection techniques can recall over half of the radio transients in the 10 per cent of the data with the highest anomaly scores. We find that the choice of anomaly detection algorithm makes a minor difference, but that feature set choice is crucial, especially when considering available resources for human inspection and/or follow-up. Active learning, where human labels are given for just 2 per cent of the data, improves recall by up to 20 percentage points, depending on the combination of features and model used. The best-performing results produce a factor of 5 times fewer sources requiring vetting by experts. This is the first effort to apply anomaly detection techniques to finding radio transients and shows great promise for application to other data sets, and as a real-time transient detection system for upcoming large surveys.

Key words: surveys – software: data analysis – software: machine learning – radio continuum: general – radio continuum: transients.

1 INTRODUCTION

The Square Kilometre Array (SKA; Braun et al. 2019) promises to deliver unprecedented data across many areas of astrophysics. Already, SKA precursors such as ASKAP (Hotan et al. 2021) and MeerKAT (Jonas & MeerKAT Team 2016) are providing a unique combination of highly sensitive observations taken at regular cadence over a wide field-of-view. These observations then allow for untargeted searches for radio transients in surveys such as ThunderKAT (Fender et al. 2016) and VAST (Variables and Slow Transients; Murphy et al. 2013). ThunderKAT observed known radio transients (e.g. X-ray binaries; XRBs), but the unique capabilities of MeerKAT have allowed for serendipitous discovery of magnetically active and flaring stars (Driessen et al. 2020, 2022a; Andersson et al. 2022; Chastain et al. 2023, 2024; Peters 2023; Fijma et al. 2024; Smirnov et al. 2025, as well as Mlangeni et al. and Nyamai et al., in preparation), scintillating active galactic nuclei (AGN) and

★ E-mail: alexander.andersson@physics.ox.ac.uk

© 2025 The Author(s).
Published by Oxford University Press on behalf of Royal Astronomical Society. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/mnras/article/538/3/1397/8042556 by Universiteit van Amsterdam user on 08 April 2026

unique pulsar behaviour (Rowlinson et al. 2022; Driessen et al. 2022b). The VAST survey's wide-field imaging with ASKAP has discovered large, scintillation-inducing plasma arcs (Wang et al. 2021a), mysterious transients near the Galactic centre (Wang et al. 2021b, 2022), and many highly variable systems on minute time-scales (Wang et al. 2023). Furthermore, recent observations at time-scales of a few minutes have led to uniquely long-period transients[1], pulsars, and white dwarf systems (Caleb et al. 2022, 2024; Hurley-Walker et al. 2022, 2023; Pelisoli et al. 2023; de Ruiter et al. 2024; Hurley-Walker et al. 2024). Finally, imaging on very short time-scales can immediately localize fast radio bursts (FRBs) to high precision (Rhodes et al. 2023; Driessen et al. 2024). The combination of all aspects of a data set – its spectral coverage, time baseline, and polarimetry – has been shown to provide a unique discovery space for pulsars and transients (Heywood 2023; Smirnov et al. 2024).

Despite this interesting zoo of sources, radio transients, and variables in observations around 1 GHz are relatively rare, of order 1–4 per cent of sources per observational field (Mooley et al. 2016; Murphy et al. 2017; Sarbadhicary et al. 2021). At the same time, current and upcoming radio facilities promise to overwhelm astronomers with data, creating a needle-in-a-haystack problem of detecting the rare and interesting transients in a deluge of high-cadence, highly sensitive observations. As a demonstration of this, current deep imaging with MeerKAT can detect ~ 6000 sources/deg^2 (Heywood et al. 2022). The SKA-Mid Array, when complete, will be able to reach the same depth in approximately 12 h.[2] If the array were to conduct non-overlapping 12 h pointings over an entire year, it would observe upwards of 4000 000 sources (over less than 1/40th of the sky). If time-variable phenomena make up approximately 2 per cent of all sources in a given radio image at any sensitivity limit, this means that the SKA-Mid will see up to $\sim 40,000$ transients and variables in its first year of operation. There is therefore a need to develop and test a range of methods in order to discover transients in data-streams, with the enormous data-rates prompting the need for low-latency searches in order to avoid storing unreasonably large quantities of data. An example of this can be seen in the case of the MeerTRAP (More TRansients And Pulsars; Stappers 2016) backend employed on MeerKAT. The MeerTRAP user supplied equipment taps into the MeerKAT data stream, commensally searching for fast transients such as FRBs and bright single pulses from pulsars. This effort requires real-time candidate processing and filtering in order to reduce the data rate to reasonable volumes, with final candidates requiring additional rapid confirmation or rejection by the team in order to not overwhelm storage capacity (Malenta et al. 2020; Jankowski et al. 2022).

This data-intensive regime is not unlike the problem that faces optical astronomers, who have spent several years preparing for the Vera C. Rubin Observatory to come online and deliver the Legacy Survey of Space and Time (LSST), consisting of millions of transient alerts per night in a cumulative volume of petabyte scale data (Ivezić et al. 2019). One proposed solution to this problem has been into anomaly detection methods – these are unsupervised (that is, untrained) machine learning models that seek to separate rare anomalies from everything else. Much of this work has been done relating to supernovae, due to their relevance to a wide range of astrophysics and cosmology but the scarcity of spectroscopic follow-up. For example, Pruzhinskaya et al. (2019) and Malanchev et al. (2021) demonstrate the applicability of isolation forests and local outlier factor algorithms (see Section 4) to detecting rare supernovae classes and novel transients. It is not only the models themselves that have been tested, but also the descriptions of the input data. As mentioned in Section 3, the choice of features – attributes or parameters that are calculated or learnt from raw data, which are then fed into downstream tasks – is paramount in machine learning. These features might be extracted based on some previously known importance (e.g. the peak magnitude of a supernova) or, for large and complex data sets, learnt from the data themselves (deep learning). For example, Villar et al. (2021) use an autoencoder framework to extract salient features from light curve simulations directly, as opposed to using previously known metrics. However, deep learning architectures are not always the answer – for example, Muthukrishna et al. (2022) demonstrate that their overflexibility can make them less suitable for detection of anomalies in real time when compared to parametric models.

One potential pitfall of these anomaly detection techniques is that anomalies identified by a specific algorithm may not match what the end-user (i.e. a science team) deem to be interesting. By incorporating domain knowledge into machine learning models to refine their performance one can, for example, separate anomalous instrumental effects from bonafide astrophysical phenomena. This is known as active learning, the core idea of which is that an algorithm can query a human annotator (sometimes called an 'oracle') to label a small subset of data from which the learner can train. If the learner can choose the data for the annotator to label, the hope is that performance will increase by labelling only a small but informative subset of the data. This active learning for anomaly detection has been tested on simulated data in preparation for LSST and real data streams from the Zwicky Transient Facility and the Open Supernova Catalog (Ishida et al. 2021; Pruzhinskaya et al. 2023).

Despite the advances in appropriate feature use, anomaly detection and active learning in optical surveys, these tools have yet to be applied to the rapidly growing field of wide-field radio transient surveys, with the most similar use-cases being for investigating anomalous spectrograms and parameter optimization in low-frequency radio data and pipelines (Rowlinson et al. 2019; Mesarcik et al. 2023). The ASTRONOMALY package (Lochner & Bassett 2021) provides a general framework for discovering anomalies in astronomical data sets, which we can use to test these different aspects of the discovery process. ASTRONOMALY has three main scientific steps consisting of feature extraction, anomaly detection, and active learning, whilst also providing data management, pre-processing, and visualization tools. Exploration of how these different stages perform, with different features and algorithms applied to a novel data set form Sections 3–4.3 of this work, though we note that they are all built into ASTRONOMALY's framework. ASTRONOMALY has been used to discover variable and flaring stars in optical surveys (Webb et al. 2020, 2021), as well as rare and unique morphologies of optical galaxies, mergers, and radio sources (Lochner et al. 2023; Etsebeth et al. 2024; Mohale & Lochner 2024).

In this paper, we describe the first exploration of applying unsupervised anomaly detection techniques to large samples of light curves from radio telescopes. This is motivated by both the ever-growing data rates from modern radio telescopes as discussed earlier, along with the need for robust methods for finding transients that improve upon previous searches. Previous studies (such as Driessen et al. 2022b; Andersson et al. 2023) have shown that many variables and transients are difficult to identify with simple statistics (see equations (2) and (3)). As we are interested in finding such transients in large data sets (e.g. from the SKA), these findings were

[1]Here, minute-time-scales are considered 'long' with respect to typical pulsar periods of < 1 s.
[2]Calculated from https://sensitivity-calculator.skao.int/mid.

Downloaded from https://academic.oup.com/mnras/article/538/3/1397/8042556 by Universiteit van Amsterdam user on 08 April 2026

part of the motivation for studying and applying anomaly detection techniques.

We make use of 8874 light curves from the MeerKAT telescope as part of commensal ThunderKAT observations, analysed by citizen scientists in search of variables and transients (Andersson et al. 2023). By its nature, unsupervised learning requires no classifications and therefore easily scales to large data sets as there is no need for costly and/or time-consuming labelling of data nor the training of large models. By the same token, however, it is hard to verify performance of unsupervised learning methods without some known 'ground truth' (see e.g. Giles & Walkowicz 2019). Therefore, we make use of the citizen science classifications as ground truth against which to compare anomaly detection methods from the ASTRONOMALY package. We describe the MeerKAT and citizen science data in Section 2 before detailing the features used in this work in Section 3. We then demonstrate how standard anomaly detection performs on these data in Section 4 and how the active learning built into ASTRONOMALY improves this performance in Section 4.3. We discuss this work and its implications before concluding in Sections 6 and 7.

2 DATA: THUNDERKAT OBSERVATIONS AND CITIZEN SCIENCE

This work builds on the data collected and presented in Andersson et al. (2023, A23 hereafter). Here, we summarize the observations and resultant data products (light curves), as well as the citizen science classifications that followed. For full details, see sections 2 and 3 of A23.

2.1 Observational data

Our observational data consist of ∼500 images taken by MeerKAT from 2018 to 2022 at 1.3 GHz, distributed across 11 fields of the sky and centred on XRBs of interest to the ThunderKAT programme. These observations, typically taken at weekly cadence, consist of scans of a primary calibrator and phase calibrator bookending a 15 min observation on-source. The observations were reduced in a standard procedure, using the `OxKAT` scripts (Heywood 2020) and their underlying packages to flag, calibrate, and image the data. The resultant set of images is heterogeneous in number of visits per field and this heterogeneity is entirely determined by the commensal nature of the observations – if an XRB faded into quiescence it was dropped from the weekly schedule. As a result some fields have only two months worth of observations, whilst fields around bright and active systems such as MAXI J1820+070, GRS 1915+105, and GX339−4 have over 50 visits.

Once imaged, light curves are extracted from the observations using the Transient Pipeline (TRAP; Swinbank et al. 2015). The TRAP extracts flux density measurements for all sources detected with an S/N ≥8 within 1.5× the primary beam radius of each observation. Each flux measurement comes from a Gaussian fit to each source, fixed to be the size of the PSF in that observation, as we are only interested in point sources and a changing PSF between epochs can induce false variability. The TraP is run such that if, in any observation, a source is found, that position is checked and force-fit in all subsequent epochs (but not in prior images), providing upper-limits in images where the source is absent (e.g. a transient). Examples of a bona fide transient, a false positive, and a stable light curve are all visible in Fig. 1, demonstrating the diversity of our data.

From running the TraP, we produce 8874 light curves in total from the 11 XRB fields. As with the images, the number of light curves per field is not uniform, with some fields having many more than others. This is primarily governed by some fields having more diffuse emission, higher noise floors and/or worse *u,v* coverage.

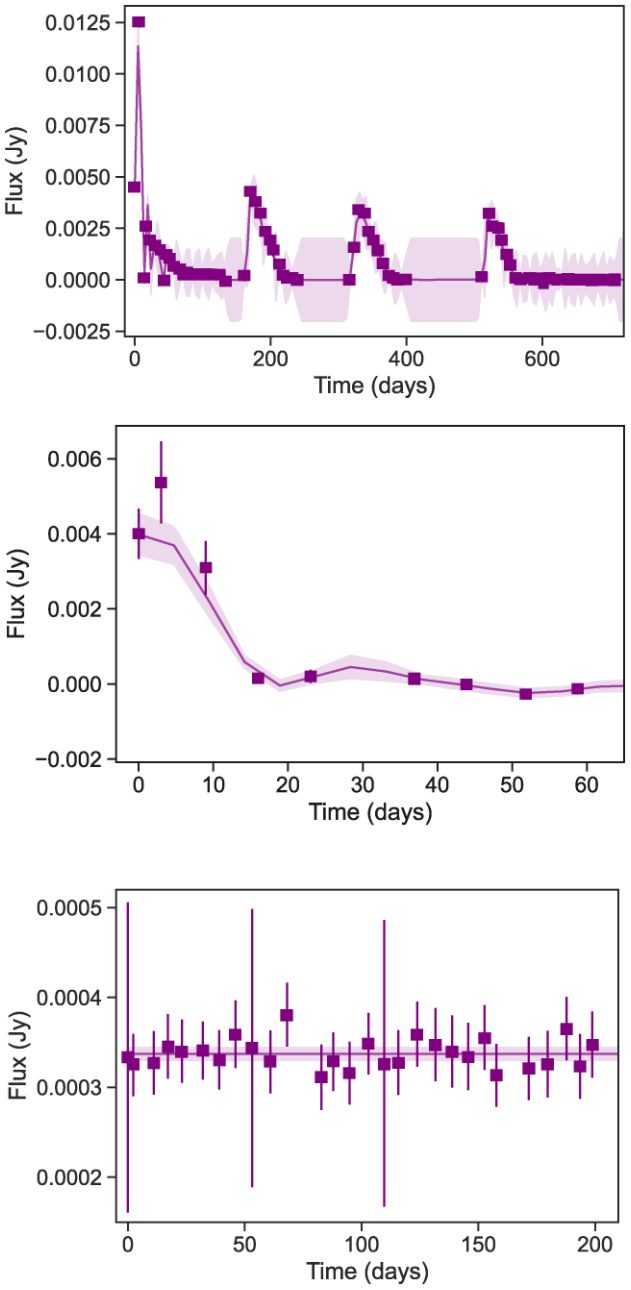


Figure 1. Example light curves of an XRB (MAXI J1820+070, top[3]), a false positive caused by bad data (middle) and a stable light curve (bottom). Our aim is for unsupervised machine learning algorithms to preferentially select light curves similar to the foremost over the latter two. The line and shaded region of each panel denote the mean and 1σ uncertainty of the computed Gaussian Process regression, detailed in Section 3.3.

[3]See Bright et al. (2020) for more on the radio behaviour of this source and Bright et al. (in preparation) for analysis of the long-term light curve.See Bright et al. (2020) for more on the radio behaviour of this source and Bright et al. (in preparation) for analysis of the long-term light curve.

Downloaded from https://academic.oup.com/mnras/article/538/3/1397/8042556 by Universiteit van Amsterdam user on 08 April 2026

2.2 Citizen science data

The 8874 light curves described in the previous section were shown to volunteers on the *Bursts from Space: MeerKAT* (BfS:MKT) citizen science project, hosted on the Zooniverse platform (A23).[4] Total of 10 volunteers look at each subject on the platform and determine whether the source in question is truly a transient/variable or a member of a different class – stable sources with flat light curves, imaging artefacts that can appear as spurious transients and so on. Once collected, these 10 votes were aggregated to give a transient vote fraction and one of us (AA) manually inspected all 381 sources where 4/10 or more volunteers said a subject was a transient or variable. From this final selection we have 168 transients and variables in our sample, consisting of 142 newly identified sources and 26 that had been previously studied.

For the purposes of this paper, we will treat the 168 variables and transients found by our volunteers and verified by us as our ground truth sample. From hereon, any reference to transients refers to this sample, representing ∼2 per cent of our light curves. Due to this large imbalance between transients and non-transients, we choose to apply anomaly detection methods in order to understand how detectable these rare systems are *without* the need for citizen science verification (e.g. in a low-latency transient detection system). As a note, we do not distinguish between variables and transients in this work. This is due to the slight observational ambiguity involved – e.g. the radio emission from a flaring star may indeed be seen in transient bursts, but the star itself is not truly a transient. The only unambiguous distinction between transients and variables is in the case of cataclysmic, one-off explosions (e.g. supernovae), but nevertheless our search methods would identify these, in addition to sources varying entirely above or around our detection thresholds.

3 FEATURE EXTRACTION

In order to make use of our anomaly detection tools, we must first calculate some features that describe our light curves. The choice of how to describe a data set is crucial in machine learning tasks and can generally be split between features based on domain knowledge or experience and deep learning, where features are learned automatically from the data. In this work, we use two domain-informed feature sets which have been used in machine learning searches for variability in previous studies. We also test the union of both these sets together, which we will refer to as the combined feature set. We choose to use these domain-informed features as opposed to those learnt from the data for several reasons. First, automatic feature extraction methods can have many thousands of parameters (internal weights and biases) to update, requiring similarly extensive data sets. With our moderate sample of ∼9000 light curves, taken from heterogeneously observed fields, we cannot expect our data to be representative of all possible underlying variability. Similarly, data-driven feature extraction can be computationally expensive, requiring many CPU- or GPU-hours to train. Finally, by using features tested in other regimes we can compare and contrast their performance between use cases. We will also compare the performance of the two feature sets to a baseline pair of parameters often used in transient detection and variable characterization (see e.g. the ThunderKAT commensal work in Section 1).

[4]https://www.zooniverse.org/projects/alex-andersson/bursts-from-space-meerkat

Every feature x used in this work is standardized by computing

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

i.e. by subtracting the mean μ and normalizing by standard deviation σ. This is a common approach in machine learning applications (e.g. Czech, Mishra & Inggs 2018) and is done so that features with large absolute ranges (or measured in different units) do not dominate. For example, if a model required computation of distances between datapoints in a 2D data set where one feature had a range $\mathcal{O}(10^4)$ whilst the other had a range $\mathcal{O}(1)$, the former would dominate the distance calculation.

3.1 Baseline features

As a baseline pair of features we make use of two parameters often used to discover radio transients. For a light curve of N flux density measurements $F_i \pm \sigma_i$, observed at frequency ν, the two variability statistics are defined as

$$\eta_\nu \equiv \chi^2_{N-1} = \frac{1}{N-1}\sum_{i=1}^{N}\frac{(F_{i,\nu} - \overline{F_\nu}^*)^2}{\sigma_i^2} \tag{2}$$

and

$$V_\nu \equiv \frac{s_\nu}{\overline{F_\nu}} = \frac{1}{\overline{F_\nu}}\sqrt{\frac{N}{N-1}(\overline{F_\nu^2} - \overline{F_\nu}^2)}, \tag{3}$$

where s, $\overline{F_\nu}$, and $\overline{F_\nu}^*$ denote standard deviation, mean, and weighted average of the flux density, respectively. Broadly speaking, η_ν, as a chi-squared test, quantifies the statistical significance of variability in a light curve, whilst V_ν describes the amplitude of variability and is sometimes known as the modulation index. We note that this two parameter setup is far simpler (lower dimensions) than the two feature sets described subsequently, which may not capture more subtle variations in the light curves. This is due to our anomaly detection techniques being designed for high dimensional data and there is no guarantee that their algorithmic definitions of anomalies match human intuition when using η_ν and V_ν.

3.2 Feets features

Next we used a feature set from the *feets* package (Cabral et al. 2018a, b) inspired by the FATS code (Nun et al. 2015). This package combines over 40 features used for time series analysis, which can be broken into approximately three subsets. First, some features calculated are simple statistical quantities, such as the mean, amplitude, and standard deviation of a light curve. The second subset of features used are included due to their known statistical or astrophysical relevance to a given class of object such as variable stars (Richards et al. 2011) and quasars (Kim et al. 2011) and include, for example, the skewness and the exponent of a light curve's structure function. Finally, several features used are derived from a Lomb–Scargle periodogram (Lomb 1976; Scargle 1982) and its resultant best fit to each light curve. We note that there are some features in this codebase that require colour information but as our data is univariate (monochromatic), we cannot employ these features in this implementation. These *feets* features have an implementation built into ASTRONOMALY's framework, making them ready to test for a given application out-of-the-box. The full table of features used is detailed in Table 1.

Downloaded from https://academic.oup.com/mnras/article/538/3/1397/8042556 by Universiteit van Amsterdam user on 08 April 2026

Table 1. Features used in this work as calculated by the *feets* package (Cabral et al. 2018a). Each light curve consists of $N = \Sigma i$ measurements of flux density $F_i \pm \sigma_i$ taken at time t_i. A similar version for the original FATS code is described by Webb et al. (2020).

Feature	Description	Inputs	Reference
Amplitude	Half the difference between the median of the maximum 5 per cent and the median of the minimum 5 per cent of fluxes	F_i	Richards et al. (2011)
Anderson Darling test	A test statistic for non-Gaussianity	F_i	Kim et al. (2009)
Auto Correlation length	Length of linear dependence of a signal with itself at two points in time, taken when the AC function reaches $1/e$	F_i	Kim et al. (2011)
Beyond1Std	Percentage of points beyond one standard deviation from the weighted mean	F_i, σ_i	Richards et al. (2011)
Con	The number of three consecutive data points that are brighter or fainter than 2σ and normalised by $N-2$	F_i	Kim et al. (2011)
η^e	$\bar{w} \sum w_i (F_{i+1} - F_i)^2 / (\sigma^2 \sum w_i)$ where $w_i = 1/(t_{i+1} - t_i)^2$	F_i, t_i	Kim et al. (2014)
Fourier amplitudes	$A_{i,j} = \sqrt{a_{i,j}^2 + b_{i,j}^2}$ for $i = 1, 2, 3$ and $j = 1, 2, 3, 4$	F_i, t_i	Debosscher et al. (2007)
Fourier phases	$\phi_{i,j} = \arctan\left(\frac{b_{i,j}}{a_{i,j}}\right) - \phi_{1,1}$ for $i = 1, 2, 3$ and $j = 1, 2, 3, 4$	F_i, t_i	Debosscher et al. (2007)
Gskew	Median-of-magnitudes based measure of the skewness	F_i	Richards et al. (2011)
Linear trend	Slope of a linear fit to the light curve	F_i, t_i	Richards et al. (2011)
Lomb–Scargle period	The period of the largest periodogram peak	F_i, t_i	Kim et al. (2011)
Lomb–Scargle η^e	η^e as applied to the phase folded light curve	F_i, t_i	Kim et al. (2011)
Lomb–Scargle R_{CS}	R_{CS} as applied to the phase folded light curve	F_i, t_i	Kim et al. (2011)
MaxSlope	Maximum absolute slope between two consecutive observations	F_i, t_i	Richards et al. (2011)
Mean $\bar{F}$	The mean flux - $1/N \ \Sigma_i F_i$	F_i	Kim et al. (2014)
Mean variance	The ratio of the standard deviation to the mean flux - $\sigma/\bar{F} \equiv V$	F_i	Kim et al. (2011)
Median Abs. Dev.	The median discrepancy of the data from the median - $\text{med}(\lvert F_i - \text{med}(F_i)\rvert)$	F_i	Richards et al. (2011)
Median buffer range %	Fraction of points with amplitude a tenth of the median flux	F_i	Richards et al. (2011)
Pair slope trend	The fraction of increasing first differences minus the fraction of decreasing first differences	F_i	Richards et al. (2011)
Percent amplitude	Largest percentage difference between either the max or min flux and the median.	F_i	Richards et al. (2011)
Q_{3-1}	The difference between the 3rd and 1st quarterlies	F_i	Kim et al. (2014)
R_{CS}	Range of cumulative sum	F_i	Richards et al. (2011)
Skewness	$\frac{N}{(N-1)(N-2)} \sum (F_i - \bar{F}/\sigma)^3$	F_i	Richards et al. (2011)
Slotted Autocorrelation Function length	Slotted auto correlation length	F_i, t_i	Protopapas et al. (2015)
Small Kurtosis	$E[(F - \bar{F}/\sigma)^4]$	F_i	Richards et al. (2011)
Standard deviation	Standard deviation of fluxes	F_i	Richards et al. (2011)
Structure Function index	The exponent of the $SF \propto \tau^\beta$ where $SF(\tau) = \langle [F(t) - F(t+\tau)]^2 \rangle$	F_i, t_i	Hughes et al. (1992)

3.3 Wavelet features

The final set of features used come from studies primarily aimed at describing and classifying supernovae (Lochner et al. 2016; Alves et al. 2022), though they have also been used in the classification of radio transients (Sooknunan et al. 2021). These features, called wavelet features, arise from the understanding that time series $f(t)$ can be decomposed into a sum of basis functions

$$f(t) = \sum_i a_i \phi_i(t) \tag{4}$$

with coefficients a_i, where $\phi_i(t)$ might for example be a set of sines and cosines (i.e. a Fourier decomposition). We can then use the coefficients of our basis functions as a set of features for later use. Here, we use the symlet family of wavelets as our basis functions (see e.g. Arfaoui, Mabrouk & Cattani 2021 for more on these), due to their successful use in previous work (Lochner et al. 2016; Alves et al. 2022). These wavelet features are calculated with minimal assumptions on the data and so it is our hope that they describe the observed light curves well, whilst by contrast the feets package's parametric calculations are made with assumptions on the statistical distributions of the data. Furthermore, wavelet transforms are translation equivariant and so the same object (e.g. a SN) observed at different times since peak brightness should produced similar coefficients.

We calculate the wavelet coefficients for our light curves using the SNMACHINE codebase developed by the LSST Dark Energy Science Collaboration (Lochner et al. 2016; Alves et al. 2022). In order to do this, we first interpolate our light curves with a Gaussian Process Regression (GP; Rasmussen & Williams 2006), allowing us to account for gaps in our uneven time series. To do this, all light curves are shifted to take place between 0 and $t_{\max}$ d, where for our data set $t_{\max} = 1300$ d, corresponding to our longest light curves in the GX339−4 field. The underlying procedure built into

Downloaded from https://academic.oup.com/mnras/article/538/3/1397/8042556 by Universiteit van Amsterdam user on 08 April 2026

SNMACHINE fits Gaussian processes from the `George` library[5], with this implementation using GPs with a mean of zero and one of two covariance kernels. The two kernels tested are an exponential square kernel of the form

$$k_{\rm exp}(t_i, t_j) = \exp\left(-\frac{|t_i - t_j|^2}{2}\right), \quad (5)$$

and a combination of $k_{\rm exp}(t_i, t_j)$ with an exponential sine squared kernel of scale parameter Γ and period P i.e.

$$k_{\rm sin}(t_i, t_j) = k_{\rm exp}(t_i, t_j) + \exp\left(-\Gamma \sin^2\left[\frac{\pi}{P}|t_i - t_j|\right]\right). \quad (6)$$

These two kernels are intended to capture any periodic behaviour whilst the exponential decays account for the idea that points separated further in time are expected to be less correlated. The kernel used is determined by minimizing the negative log-likelihood of the GP with the SciPy implementation of a limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm (Byrd et al. 1995; Virtanen et al. 2020). Once the Gaussian Process has been fit, we evaluate the regression every ∼5 d (i.e. at 276 instances), approximately matching our typical observational cadence. Example Gaussian process regression fits can be seen in Fig. 1. For light curves shorter than $t_{\rm max}$ in length, the remaining datapoints are assumed to be 0.

Once all light curves are interpolated by the regression, the wavelet transform is calculated, producing a set of 275 highly redundant coefficients. In order to reduce this high dimensionality, we run a Principal Component Analysis (PCA; Pearson 1901), which transforms the data into an orthogonal basis. In this now-orthogonal data space, we choose to preserve the 40 largest eigenvalues for each light curve, which captures 99.999995 per cent of the summative variance of the data. This results in reducing the data space from 276 features per light curve to only 40, while retaining almost all of the variance. Fewer principal components would result in faster downstream computations, however, as mentioned earlier, anomalies (transients) constitute a small minority of our sample whilst the majority are stable light curves which should require few wavelet features to describe. Furthermore, light curves shorter than $t_{\rm max}$ are assumed to be 0 past their final datapoint, which will contribute towards the majority of the data set being non-variable. As a result, initial testing with only 10 principal components showed markedly worse performance. This shows that the diversity of light curve lengths affects the resultant wavelets and their principal components retained, which may bias results. The choice of 40 principal components is a balance between reducing the dimensionality of the features whilst trying to preserving enough information such that the anomalies, which constitute a small minority of all of the light curves, are retained. In a truly unsupervised setting, it is not possible to know how many eigenvalues to preserve as one has no labels to train from.

Finally, it is worth noting that extracting this set of features is the most computationally expensive part of our analysis. However, even the slowest step, the Gaussian process regression, runs in 10 min for the entire database on a standard desktop using a 3.5 GHz quad-core processor. Also, all of these features, once calculated, can be saved and these routines then need not be run again, making static comparisons like this work fast to conduct. For a real-time system features may be re-calculated for every new datapoint, however, this may not be optimal or required. For example, the Gaussian Process Regressions that underlie the wavelet features scale naïvely as $\mathcal{O}(N^3)$ and so calculating them for each new datapoint may be prohibitive. For larger data sets or for a real-time transient detection system a balance between computational scalability and information gained may be required and could dictate what feature sets can be deployed for a given science case.

[5]https://george.readthedocs.io/en/latest/

3.4 Feature space and clustering

An obvious question to ask is how well the feature sets and baseline feature pair separate transients (anomalies) from the rest of the distribution. For a perfect feature set, we would want all transients to be separated cleanly from all non-anomalies. To inspect this we make use of a Uniform Manifold Approximation and Projection (UMAP; McInnes, Healy & Melville 2018), a technique for dimensionality reduction. Briefly, the UMAP algorithm works by building a topological representation of data in its N-dimensional space (where N is 45 and 40 for the *feets* and wavelet features, respectively). Once this has been constructed, the UMAP algorithm works to optimize a low dimensional (e.g. 2D) representation of this topological space as measured by cross entropy. Datapoints that are similar in the N-D space should then remain nearby in the 2D representation, which can then be used for visualization or further downstream tasks.

As a baseline, the feature space of (η, V) is already in 2D so there is no need to reduce its dimensionality. This feature space is shown in Fig. 2. This figure demonstrates that many transients and variables are inliers, as found in A23. The presence of systematics in some fields (e.g. MAXI J1820) can be seen, along with the difficulty in separating our transients (anomalies) from all background sources. This was one of the main motivations for testing the subsequent feature sets for anomaly detection.

The *feets* feature space for our data set can be seen in Fig. 3, where stars are our citizen-scientist verified anomalies (transients) and the colourbar indicates which observational field a given light curve is from (see Section 2.1). We can see that sources in a given field are clustered together. This is likely due to observational systematics varying on a field-by-field basis. For example, the SAX J1808 field only received 6 observations and, resultingly, exists in a very different area of parameter space to other fields. Whilst this might be of slight concern, we do not use these 2D representations for downstream tasks, but rather as sanity checks and in order to judge qualitatively how well our data (anomalies/transients and everything else) are described by a given feature set. Moreover, for a real-time or low-latency detection system, what is of more importance is how well transients separate from background data *within* a given field – as in such a system, it is only against sources in the same observation that we would compare. Indeed, we can see that anomalies are somewhat clustered and that, whilst they tend to lie in their respective field of light curves, they are often on the edges of each region. A demonstration of this can be seen for the GX339−4 field shown in the inset of Fig. 3.

Similarly, the feature space of our wavelet coefficients in Fig. 4 shows field-dependent clustering. However, in this case, there are clear strings of points in feature space that correspond to the observational fields with longer light curves (e.g. J1858, GX339−4). The majority of these sources are not transient and so perhaps these strings of points arise from requiring the same number of coefficients, of similar numerical values, to reproduce similar light curves. By contrast, shorter light curves and variable sources are clustered in the same area of parameter space (inset), however, as are some non-anomalous background sources. This distinct cluster of data points may give insight into the performance of the wavelet features in ASTRONOMALY's models in Sections 4 and 4.3.

Downloaded from https://academic.oup.com/mnras/article/538/3/1397/8042556 by Universiteit van Amsterdam user on 08 April 2026

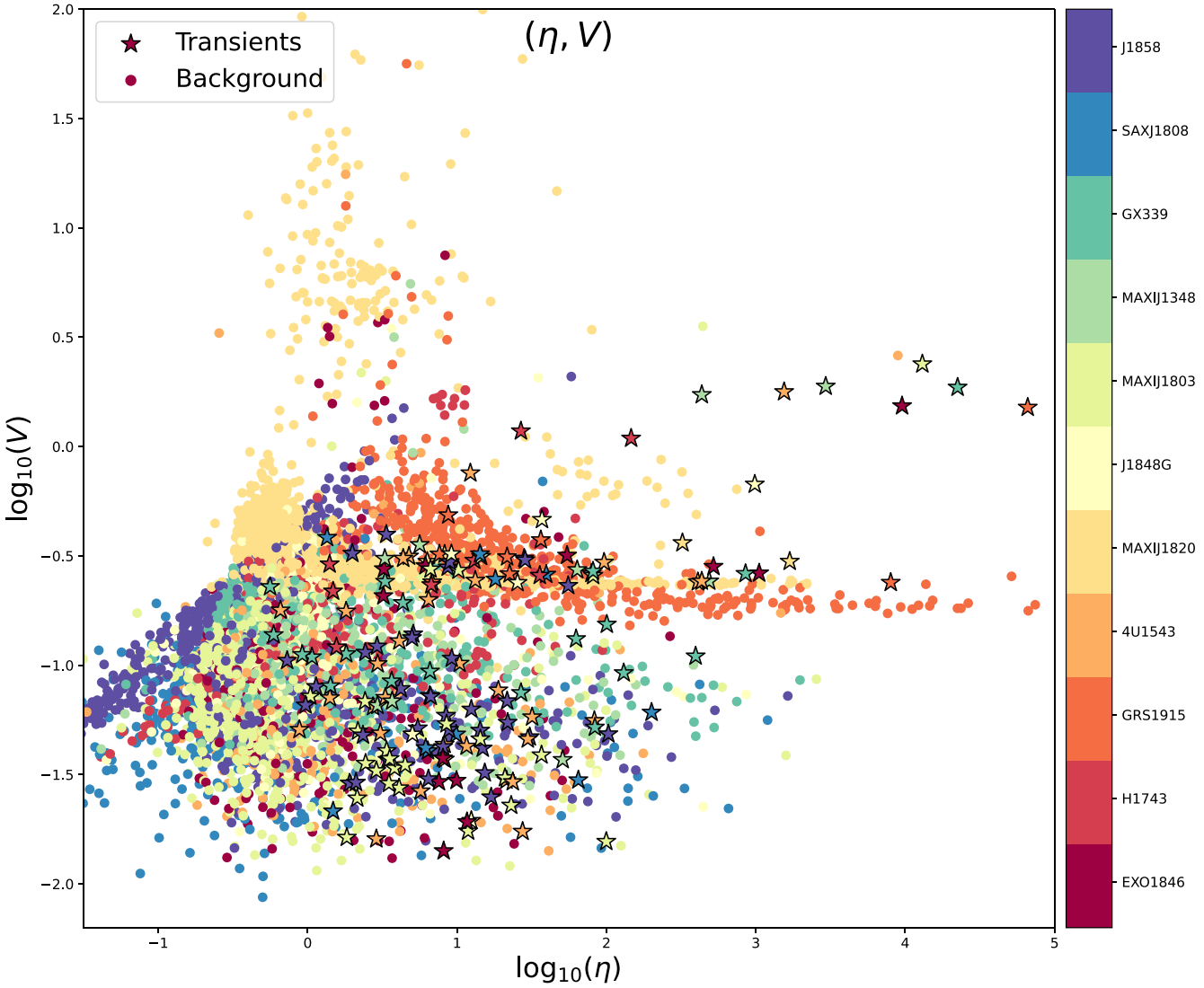


Figure 2. The (η, V) feature space, colour coded by observational field. Stars indicate volunteer-verified anomalies (transients). There are clear field-dependent systematics affecting, for example, the MAXI J1820+070 field.

4 ANOMALY DETECTION AND ACTIVE LEARNING

Having calculated features for our light curves, we move on to the core anomaly detection step of ASTRONOMALY. To do this we pass our standardized features into two commonly used anomaly detection algorithms, detailed below, as implemented in the `scikit-learn` package (Pedregosa et al. 2011). Anomaly detection is typically done by each algorithm having a definition of what is normal and assigning subjects that are far in parameter space from normal as highly anomalous, without any user labels or training. Once the degree to which something is anomalous has been ranked by a given algorithm, ASTRONOMALY scales this to a human reading of 0 to 5, where 5 is most anomalous. The output algorithm scores are linearly rescaled such that 5 represents the most anomalous source while 0 is the least.

By its nature, unsupervised learning has no pre-defined metric to calculate success. However, as noted earlier, we have votes from citizen scientists for all of our light curves from BfS:MKT, which in turn have been verified by us. Throughout this work, we will assume that these verified volunteer votes are an absolute ground truth against which to compare our anomaly detection methods. We can then use the positions of our citizen-science verified transients in our anomaly score-ordered list to calculate metrics of success – the more transients that are ranked as highly anomalous, the better. It is important to point out that ASTRONOMALY never decides whether or not something is anomalous and simply orders the entire list by anomaly score. The use of information from our volunteers is what allows us to then calculate metrics of success such as recall and precision (see Section 4.2). A contrasting method for cases without such citizen science data might be to define some threshold in an anomaly score-ordered list above which things are said to be anomalies, providing the binary classes on which to calculate performance.

4.1 Anomaly detection algorithms

The first of the two anomaly detection methods used herein is the Local Outlier Factor (LOF; Breuniq et al. 2000). The LOF is a density-based method that measures the deviation of the density of a given sample with respect to some user-defined number of nearest

Downloaded from https://academic.oup.com/mnras/article/538/3/1397/8042556 by Universiteit van Amsterdam user on 08 April 2026

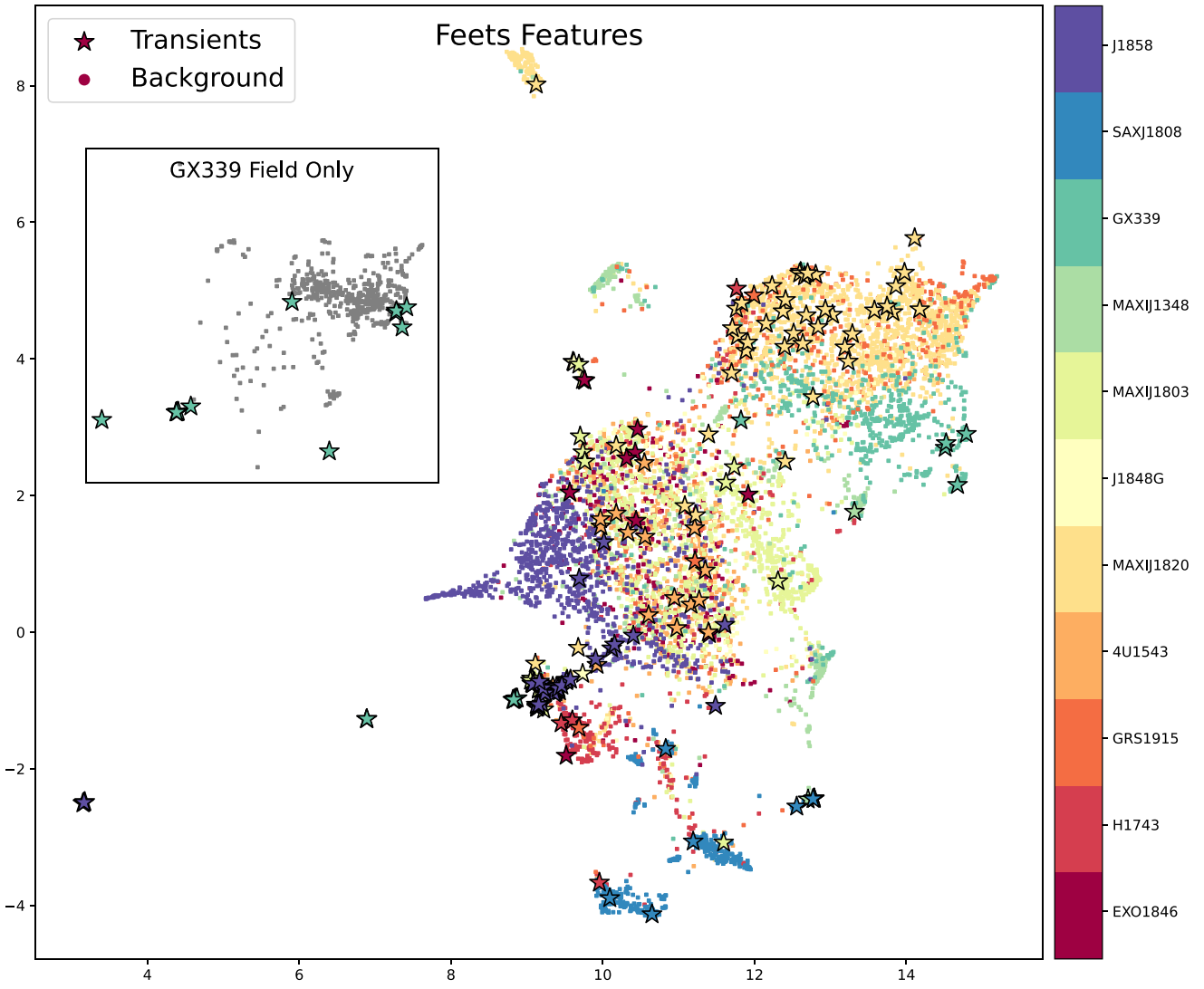


Figure 3. The UMAP of our *feets* feature space, colourcoded by observational field. We see that clustering is mainly a function of observational field, aligning with the heterogeneous nature of the data set. The axes are unlabelled as these numbers are arbitrary combinations of many features and have no physical meaning. The inset is a demonstration of how the data of one particular field (GX339) is distributed, as mentioned in the text.

neighbours, k. We set $k = 100$ throughout this work, approximately 1 per cent of our sample. Values of k between 20 (the default) and 200 were tested, producing changes in performance at the sub per cent level. As noted earlier, without labelled data it would be difficult to optimize this (and other) hyperparameters. The LOF is local in that the resultant anomaly score depends on how isolated the object is with respect to the surrounding neighbourhood. More specifically, the LOF of datum A is a measure of the average local density of all k neighbours divided by A's own local density. One advantage of this local approach is that it might pick up anomalies missed by global density measurements. That is, a point only a small distance from a very dense cluster might be an anomaly, whereas a point in a sparse cluster might still have the same density as its neighbours and so not be an outlier despite its low absolute density value. For more on the distinction between global and local anomalies, see the discussion in Rogers et al. (2024).

The second method used is Isolation Forest (IF; Liu, Ting & Zhou 2008). IFs take inspiration from the supervised random forest algorithm (RF; Breiman 2001) and operate on the core principle that anomalies are both few in number and different compared to normal data. In an IF an ensemble of decision trees randomly split random features, growing each 'branch' in the tree until all considered data are isolated. Given that anomalies should be few and different, they are easier to isolate than normal data that are close to each other in feature space. Therefore data with shorter decision branches are more likely to be anomalous, whilst those that take many splits to isolate are more similar to the rest of the sample. The final anomaly score is an average score over all trees in the forest. This model is particularly applicable to high dimensional data as there is no costly calculation of distances and therefore has a linear time complexity (i.e. twice the input data takes twice the runtime), c.f. LOF which can take up to $\mathcal{O}(N^2)$ (Breuniq et al. 2000).

4.2 Anomaly recall and precision

In order to assess the performance of our models we can ask out of all the anomalies that exist, what fraction did we find in the top n sources? This quantity is known as the recall (also referred to as the completeness, true positive rate,or the sensitivity) and is only calculable with ground truth labels for an entire data set, as without

Downloaded from https://academic.oup.com/mnras/article/538/3/1397/8042556 by Universiteit van Amsterdam user on 08 April 2026

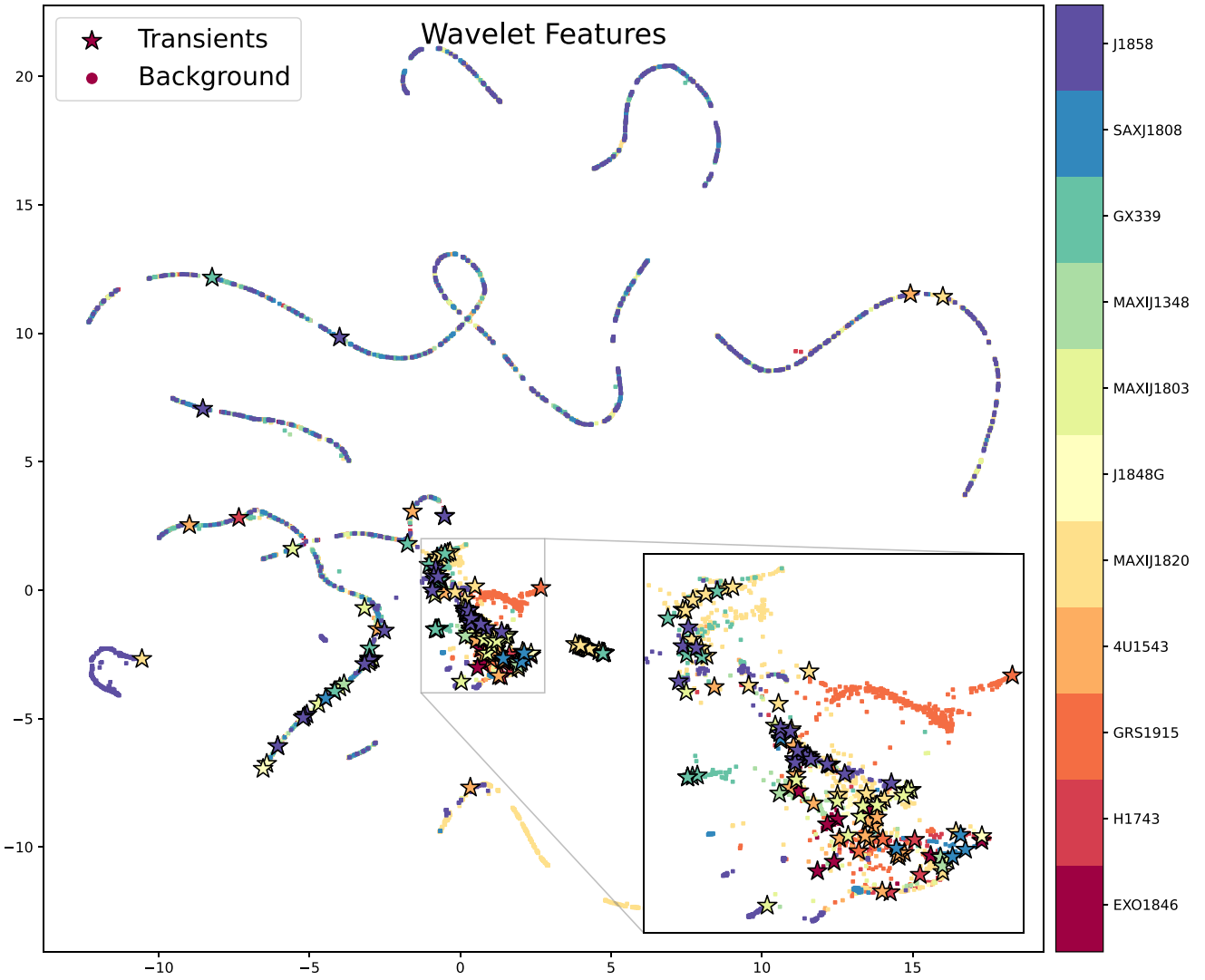


Figure 4. The UMAP of our wavelet coefficient feature space, colour coded and labelled as in Fig. 3. In this instance, the observational fields with longer light curves (e.g. J1858, GX339−4) show distinct patterns, whilst most other sources are in a very dense region of parameter space, as shown by the inset zoom.

this it is impossible to know how many anomalies are missed (false negatives). Recall is defined as

$$R = \frac{\text{True Positives}}{\text{True Positives} \ + \ \text{False Negatives}}, \tag{7}$$

where for us the denominator is our total number of transients (anomalies) in the data, i.e. 168. This recall value, whilst simple, corresponds well with preferable performance consisting of many more anomalies high in the list for the user to see quickly. We can calculate the recall R as a function of n until we reach the end of our data set ($n = 8874$), by which time we will have recovered all sources in the list. We define R_{10} and R_{30} as the recall of transients in the top 10 and 30 per cent of our list, respectively, which we will use as quantitative measures of performance throughout this work. We stress again that it is impossible to calculate this recall without labels for all transients, as without this we would not know how many transients fall below a given threshold i.e. the rate of false negatives. As such, looking at the top 10 or 30 per cent of the anomaly score-ordered lists is somewhat arbitrary and depends on, for example, the resources available for follow up or further citizen science classification.

We can also ask what fraction of the top n sources are transients. This is known as the precision, also referred to in astronomical literature as the purity, defined as

$$P = \frac{\text{True Positives}}{\text{True Positives} \ + \ \text{False Positives}}. \tag{8}$$

This can be thought of as a measure of the hit-rate of our anomalies. Namely, how many transients (anomalies) would there be in the top 10 or 30percnt of the data? By definition, as n approaches the total length of our data set (8874 sources), the precision will tend towards the aggregate rate of transients in our data i.e. ~ 2 per cent. precision, unlike recall, only depends on the labels of sources above some cut-off point set by the user and so is calculable even for a largely unlabelled data set.

Downloaded from https://academic.oup.com/mnras/article/538/3/1397/8042556 by Universiteit van Amsterdam user on 08 April 2026

Figure 5. The ASTRONOMALY GUI showing subjects in anomaly-order. Using this interface, we can easily inspect our data and provide labels on some subset of objects in order to perform active learning.

4.3 Active learning

Once anomaly scores have been calculated for all data, ASTRONOMALY allows for interactive visualization by use of a GUI, seen in Fig. 5. This consists of sources listed ordered by anomaly score, displaying the raw data, its features and any associated metadata. From this interface we can label sources on a range of 0–5. This allows for active learning, whereby an additional model balances the raw anomaly score with user input and calculates a trained anomaly score. This then acts as a recommendation engine, displaying sources with higher trained anomaly scores preferentially to the user, speeding up anomaly detection.

The core of this active learning process is described by ranking objects with a new score

$$\hat{S} = S\tanh(\delta - 1 + \operatorname{arctanh}(\tilde{U})), \tag{9}$$

where S is the default anomaly score, δ represents a distance penalty term, and $\tilde{U}$ describes a relevance score given by

$$\tilde{U} = \epsilon_1 + \epsilon_2 \left(\frac{U}{U_{\max}}\right). \tag{10}$$

Here, ϵ are normalization constants and the user input score is U, normalized by $U_{\max} = 5$. Since U will only ever be input by the user for a small fraction of a data set, its value for all remaining sources is estimated by a random forest regression calculated with 100 estimators. This regression will have uncertainty, quantified by δ, which is large when an object is far from a human-labelled neighbour and small when near to labelled data. In summary this means that, in regions far from human-labelled data ($\delta \gg 1$), $\hat{S}$ reverts to its original score as $\tanh \to 1$. Similarly, if $U \sim U_{\max}$ then S will not change due to high human ranking. Conversely, regions of feature space near labelled data ($\delta \ll 1$), with low user scores will be downweighted, corresponding to relegating uninteresting data to lower in our anomaly-ranked list.

In this work, we actively label the top ∼2 per cent of each feature-model pair, which takes of order 10 min by an individual (AA), applying a consistent methodology throughout. That is, examples of clear variability or known transients were ranked highly ($U = 4$ or 5), whilst low-significance light curves, or those showing clear false positive behaviour – e.g. where one epoch shows large changes in flux due to a different calibrator or worse image quality – were ranked at 1 or 0.

The rate of retraining in ASTRONOMALY is entirely determined by the user(s). Any number of sources can be labelled between training iterations, with between 10 and 20 sources proving to be a qualitatively good balance between providing enough data for training whilst updating regularly enough for efficient labelling. As

Downloaded from https://academic.oup.com/mnras/article/538/3/1397/8042556 by Universiteit van Amsterdam user on 08 April 2026

an example of this procedure, if the user labelled a few sources that were deemed uninteresting in the same way this would, upon retraining, down-weight further instances of similar light curves (that is, ones with similar features). However, it is worth noting that this labelling is inherently subjective and dependent on a user's end goal. An instrument scientist interested in flux calibration would score sources differently to an astronomer interested in stellar radio flares. Active learning attempts to quantify this and balance an algorithmic definition of anomalous with a user-dependent idea about what is an *interesting* anomaly. Finally, we note that the regressor used to model user interest is by default a RF but that using a Gaussian process provides an improvement on this for some data sets, which has been incorporated into the most recent versions of ASTRONOMALY (Walmsley et al. 2022b; Lochner & Rudnick 2025).

5 RESULTS

5.1 Baseline anomaly detection

The recall curves for both anomaly detection algorithms applied to all four feature sets can be seen in Fig. 6. These are normalized on both axes i.e. $x = 0.1$ corresponds to the top 10 per cent most anomalous sources and $y = 0.5$ corresponds to half of the bona fide 168 variables found by volunteers. Better performance corresponds to curves with higher gradient at low n, appearing pushed up into the upper left corner. It can be clearly seen that (η, V) show poor performance on this data set, essentially matching random chance, in line with A23's findings. As noted earlier, this will be due to a combination of applying a different methodology to the intended use of η and V and the inherent biases of these two statistics – e.g. η is correlated with signal-to-noise squared (see equation 2). Furthermore, our ground truth data consist of sources chosen irrespective of their (η, V) values. Contrastingly, we see that our main feature sets are able to recover a majority of our volunteer-verified transients (anomalies) in a minority of the data. This is the first demonstration of anomaly detection as applied to radio time series.

When comparing performance between anomaly detection models (i.e. for the same feature set), we see minimal variation and qualitatively similar behaviour between recall curves. For the *feets* and wavelet features, we see 'broken power-law' type behaviour, with an approximately constant gradient initially, followed by a knee and then worse performance. When combining these feature sets, performance is similar, though the IF and LOF diverge more after $x = 0.1$ before reconverging. Table 2 lists all R_{10} and R_{30} values for feature-model set pairs. The largest discrepancies between models occur at the 'top' of our list, with R_{10} values varying by 0.1 for *feets* features and 0.14 for wavelets. By contrast, R_{30} values differ from 0.03 to 0.07, indicating that model choice most affects the highly anomalous sources in our data set. When combining the *feets* and wavelet features, this difference between R_{10} and R_{30} values is less pronounced. It is worth noting that several iterations of each feature-model pair were tested, with performance never varying by more than $\sim$1 per cent. As a result, all values are rounded to this precision.

Differences in feature set have a much greater effect than changing the anomaly detection method – for the same model R_{10} changes by factors of $\sim$3.5 between best- and worst- performing feature sets. Similarly, R_{30} for the baseline parameters reaches only $\sim$0.34, whilst the wavelet features can recover up to 91 per cent of volunteer-verified transients. This underscores the importance of using features that best describe a given data set.

We can also calculate an aggregate recall performance by simply integrating the AUC of each feature-model pair's recall curve. The corresponding AUC values for each feature-model instance are listed in the legend of Fig. 6. As already mentioned, differences between feature sets are greater than differences between anomaly detection models. These AUC values show that the wavelet-IF combination has the best recall performance at an AUC of 0.83. However, this is a summary over all the data and does not consider the actual finding of transients in a live setting. By this we mean that one must consider how many anomalies a human will be able to find in a reasonable amount of time. As discussed above and can be seen in the lower panel of Fig. 6, if a user can only inspect the top 10 per cent of the data set, then the feets–LOF analysis performs the best, with $R_{10} = 0.52$. However, if there are resources available to inspect 30 per cent of the data set (e.g. computation, citizen scientists, and a team of experts) then the LOF performance on wavelet features recalls 91 per cent of transients, 7 per cent better than the next-best feature-model pair ($R_{30} = 0.84$).

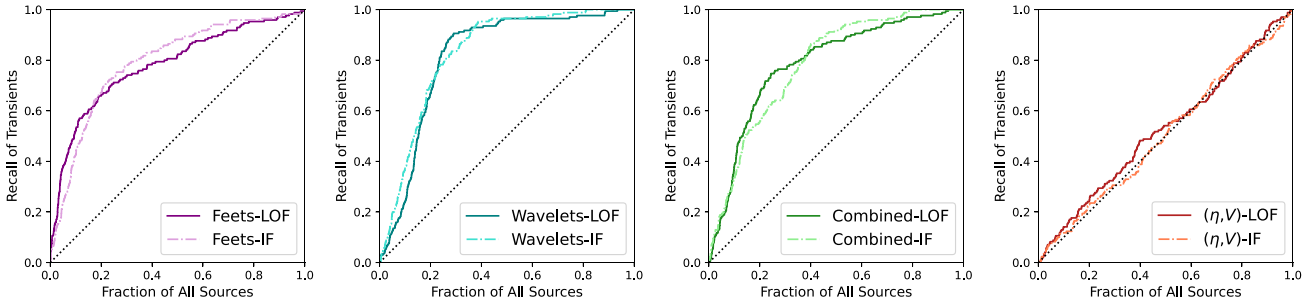
The precision values for all feature-model pairs can be found in Table 2, again looking at 10 and 30 per cent of our anomaly score-ordered lists. For the *feets* and wavelet feature sets, precision is always at least a factor of 2 better than the corresponding baseline performance. In general, the same trends hold as with our recall values – differences between feature sets are greater than differences between anomaly detection models. Likewise, we can see that the *feets* features produce the best results when looking at the top 10 per cent of the data, with a hit-rate of over 1 in 10, compared to the baseline of 1 in 50. The results 30 per cent of the way through the anomaly score-ordered lists, as seen with the recall values, show that the wavelet-LOF instance performs the best, but only marginally. Nevertheless, we can see that for either feature set or their combination, as applied to both of our anomaly detection algorithms, the hit-rate of anomalies in the data is always a factor of 2 or greater when compared to our baseline. This directly translates into the number of sources that would need to be sifted out before finding a new transient.

In summary, the choice of model used makes a small difference, though LOF is marginally preferred. When considering feature sets, the baseline features perform the worst, recovering transients almost uniformly through the anomaly score-ordered lists. The *feets* features perform the best at the beginning of each anomaly score-ordered list, recovering 52 per cent of transients in 10 per cent of the data and increasing the precision above random by a factor of 5. However, the wavelet features perform better when considering a larger fraction of each anomaly score-ordered list. Using the wavelet features over 90 per cent of anomalies are recovered in 30 per cent of the data, at a precision level that is three times higher than the baseline features.

5.2 Active learning improvement

To quantify how much active learning changes our results, we subtract each updated recall curve from those in Section 5.1. The residual curves for all feature-model pairs can be seen in Fig. 7. We can immediately see that the wavelet feature instances improve increasing recall by as much as 10 per cent within the most anomalous region of feature space. Active learning applied to the *feets* features improves anomaly recall for the most anomalous sources, but once $\sim$15 per cent of the data have been analysed, minimal improvement is seen. This is reflected in the respective R_{10} and R_{30} values seen in Table 2, which remain almost unchanged to two decimal places. The combined feature set shows a dramatic increase in performance with the LOF algorithm, by up to 20 percentage points at times. This corresponds to recovering approximately half of the transients in 8 per cent of the data compared to only around one third beforehand.

Downloaded from https://academic.oup.com/mnras/article/538/3/1397/8042556 by Universiteit van Amsterdam user on 08 April 2026



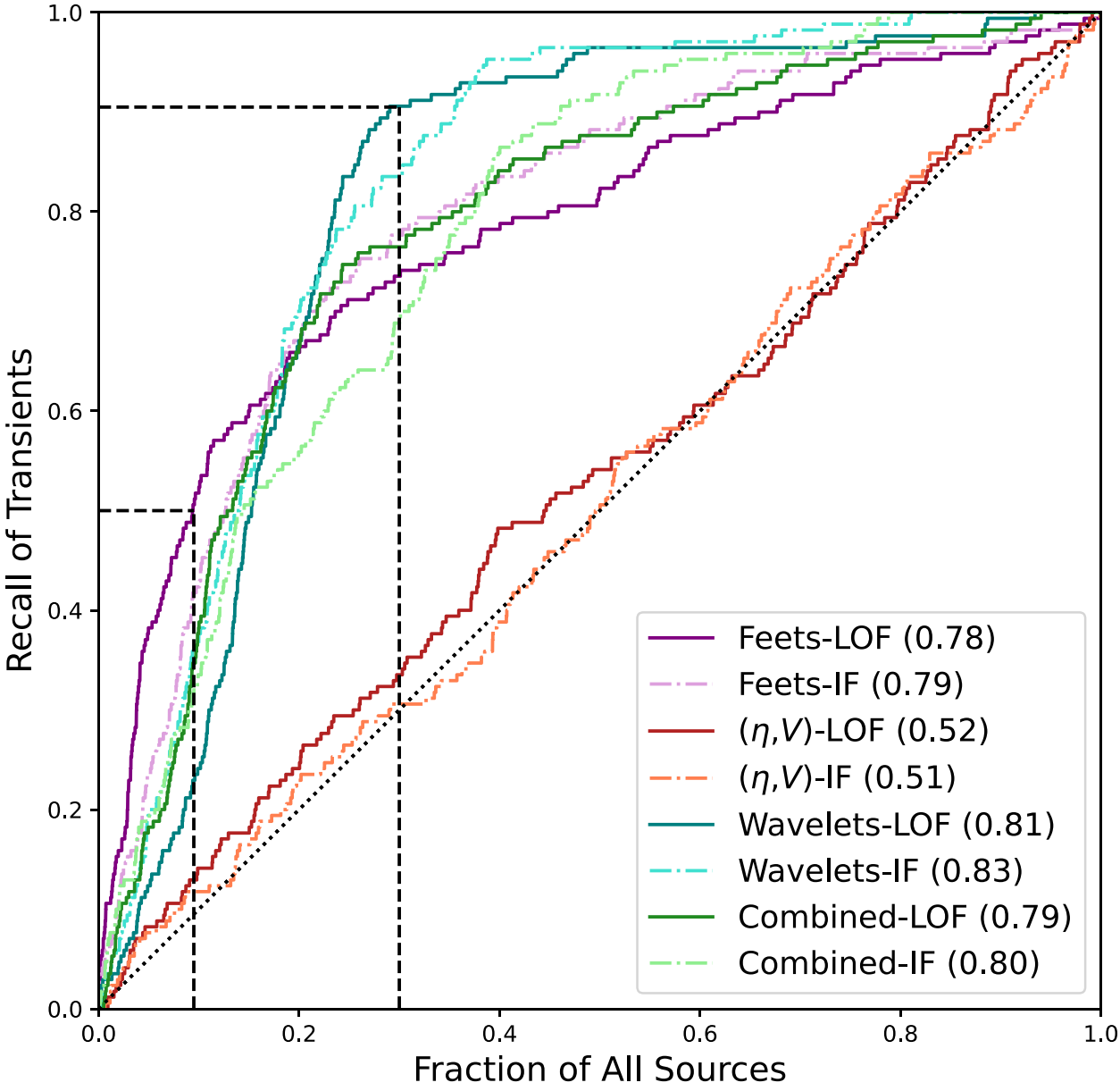


Figure 6. Upper panel: The recall curves for both anomaly detection algorithms as applied to four groups of features derived from our radio light curves. The data are ordered according to the anomaly score from each algorithm, where for example $x = 0.3$ corresponds to the most anomalous 30 per cent of the data set. The dotted line in each corresponds to $y = x$ i.e. random chance. Lower panel: A combined plot of all eight recall lines for better comparison between feature sets. Dashed lines make comparisons (see text) for recall rates at 10 and 30 per cent of the data set. The values in the legend correspond to the area under each recall curve (AUC).

Downloaded from https://academic.oup.com/mnras/article/538/3/1397/8042556 by Universiteit van Amsterdam user on 08 April 2026

Table 2. Specific recall (R) and precision (P) values, rounded to two decimal places, for each feature-model pair, measured after ten and thirty per cent of the most anomalous sources have been 'seen'. Each pair of values listed are the default, **active** iteration of that feature-model instance.

	Feets		Wavelets		(η, V)		Combined	
	LOF	IF	LOF	IF	LOF	IF	LOF	IF
R_{10}	0.52, **0.52**	0.42, **0.45**	0.24, **0.32**	0.38, **0.42**	0.14, **0.16**	0.12, **0.11**	0.38, **0.51**	0.34, **0.37**
R_{30}	0.74, **0.73**	0.78, **0.78**	0.91, **0.91**	0.84, **0.84**	0.34, **0.32**	0.31, **0.30**	0.76, **0.78**	0.69, **0.74**
P_{10}	0.11, **0.11**	0.09, **0.10**	0.05, **0.06**	0.07, **0.08**	0.03, **0.03**	0.02, **0.02**	0.07, **0.10**	0.07, **0.07**
P_{30}	0.05, **0.05**	0.05, **0.05**	0.06, **0.06**	0.05, **0.05**	0.02, **0.02**	0.02, **0.02**	0.05, **0.05**	0.04, **0.05**

Figure 7. Changes in performance due to active learning for each feature-model pair, where $y = 0$ indicates no improvement. The x-axis refers to how far down each anomaly score-ordered list is being considered. This plot has been truncated at $x = 0.4$ as, after this point, all curves vary minimally about 0, with two figures shown for clarity when comparing between curves. It can be seen that, when using the wavelet features, improvements in recall of up to 12 percentage points can be seen for both anomaly detection algorithms, whilst the *feets* features improve only moderately. When combining *feets* and wavelet features, there is a 20 percentage point increase when using the LOF, whilst the IF improvement is similar to that of each feature set alone.

By contrast, the combined feature set as applied to the IF shows approximately the same qualitative improvements as when applying the feature sets alone. As seen in the previous section, (η, V) show no improvement and in some cases perform worse than before. This is in part because our active learning procedure involves labelling the sources at the top of our anomaly score-ordered list. Given that the anomaly detection performance using the standard (η, V) feature pair essentially matches random chance and that transients make up $\sim$2 per cent of our sample, labelling only 2 per cent of the data will not provide the regressor with many positive anomalies from which to learn.

The precision values of Table 2 show minimal improvement following active learning. Only the P_{10} values improve by, at most, 1 per cent, whilst all P_{30} values remain the same to two decimal places. This is in part due to the asymptotic behaviour of these precision calculations as mentioned earlier, namely that as n reaches our total sample size, the precision must approach the overall rate of transients in our data of ~ 2 per cent. As a result, it is hard to improve our P_{30} values much when there are so few anomalies to find in our data set.

In terms of AUC values, we see minimal change compared to the default anomaly detection procedure, with all *feets* and (η, V) values

Downloaded from https://academic.oup.com/mnras/article/538/3/1397/8042556 by Universiteit van Amsterdam user on 08 April 2026

varying at the sub-per cent level and wavelet instances improving by only 0.01. This matches what we see in Fig. 7 where after $x = 0.3$ all curves vary only minimally around no change ($y = 0$). As before we stress that this metric does not match the user-driven experience of finding anomalies preferentially in as small a data volume as possible. As before, the feets-LOF iteration provide the most transients in 10 per cent of the data, though the wavelet instances perform almost as well following the active learning.

6 DISCUSSION

6.1 Feature performance and comparison

In this work, we have demonstrated the first application of anomaly detection to the task of finding transients with radio interferometers. In general, we see that appropriate features, models, and domain knowledge via active learning allow us to recover over half of our volunteer-verified transients (anomalies) in one-tenth of the total data volume. Similarly, we have shown that the number of sources a human must inspect in order to find transients can as reduced by a factor of 5 compared to baseline performance. The choice of model used makes a small but noticeable difference, with the LOF algorithm generally performing better. The choice of feature or representation is a much stronger determinant of which and how many anomalies can be recovered. For example, when dealing with the large and heterogeneous sample of data used in this work, it is clear that a simple 2-parameter setup fails to represent the data in a way that allows the subsequent anomaly detection models to succeed. This is not a surprise, given that η and V are intended for intra-field comparisons i.e. between light curves in a single set of observations. Indeed these are still useful statistics that have aided in the discovery in numerous transients from the ThunderKAT observations (see Section 1).

On the other hand, the *feets* features show good promise at recovering transients and can recall over half of our anomalies in less than 10 per cent of the data. Our wavelet features, after active learning, perform almost as well as the *feets* features with respect to R_{10} and are by a significant margin the best feature set when comparing R_{30} values. The large improvement seen in the wavelet features may be related to the clear clustering of anomalies seen in Fig. 4, where the majority of our transients are located in a relatively small part of this representation. Therefore we might conclude that if active learning adequately highlights sources from this region of feature space, performance will increase.

The combined feature set performs similar to both individual sets. The recall plot of the combined features is effectively an average of the *feets* and wavelet feature sets. This implies that while both feature sets contribute information, they are not entirely orthogonal. The additional features will make some sources appear more anomalous while others will now have more in common with the larger population. The base anomaly detection algorithms are not able to make effective use of the additional features. However, applying active learning with the LOF dramatically improves performance of these features, improving recall by up to 20 percentage points. By using this training data, active learning can far more effectively take advantage of the additional information provided, leveraging either the *feets* features or the wavelets depending on the region of feature space in which the labelled source is found. To avoid these differences caused by the initial anomaly detection algorithm, future active learning procedures may skip this step entirely (Lochner & Rudnick 2025).

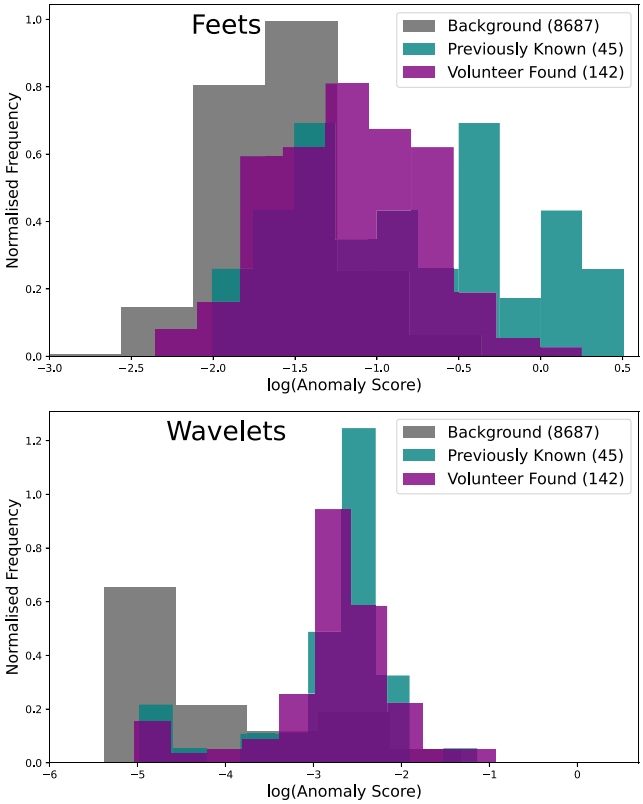


Figure 8. The distributions of the output scores of LOF with active learning, as applied to the *feets* features (upper) and wavelet features (lower), with histograms normalized by area. Volunteer found sources are those uniquely discovered by citizen scientists in A23, whilst those labelled as previously known are those that had been already identified by other works. All 8687 remaining sources are labelled as background.

A comparison between the *feets* and wavelet features can be seen in Fig. 8, where the distributions of actively learnt scores between different classes are shown. In both instances, background (i.e. non-anomalous) sources receive lower anomaly scoures as expected. However, when comparing between previously known transients (45 sources, see A23 for the listed XRBs and other commensal discoveries) and those found by volunteers (142 sources), it can be seen that when using the wavelet features, these two classes produce similar, tightly peaked distributions in score i.e. they are approximately equally anomalous. However, when using the *feets* features, the distribution of both desired classes are broader, with volunteer-found sources overlapping slightly more with background objects. This implies that, if searching only for sources such as the previously known XRBs, the *feets* features may be the feature set to use, but that the wavelet features can find both classes of transient with equal aptitude.

We note that one could spend a great deal longer curating the perfect feature set for a given task, and there is no guarantee that the features tested herein form the optimal choices. For example, we know that the heterogeneity of our time series and their sampling, along with the imputation of shorter light curves, likely induces biases in the wavelet calculations. However, without labels, it would not be possible to tune hyperparameters in any stage of the process. Nevertheless, that pre-existing methods work well in this regime is a significant step towards finding radio transients in real-time surveys. Future work may investigate whether, for example, autoencoder architectures (e.g. Villar et al. 2021) provide

Downloaded from https://academic.oup.com/mnras/article/538/3/1397/8042556 by Universiteit van Amsterdam user on 08 April 2026

a way to extract meaningful features directly from the radio time series.

As a final point on feature selection, we have used Gaussian Processes only for interpolating light curves, though they serve as powerful tools in their own right. When fitting a GP, the covariance kernels used have hyperparameters that can provide alternative metrics for detecting variables and transients. This idea is explored further in Fu et al. (in preparation), who use a multiterm covariance kernel and its hyperparameters to find candidate transients in the same data used here. This study finds that the joint distribution of hyperparameters is a good discriminant for separating variable and transient candidates from stable ones and goes on to detect transients missed by traditional metrics and by A23.

6.2 Computational scalability

With regards to scalability, all processing has been performed on a standard CPU on a desktop with no attempt at optimization. Despite this, all feature extraction and anomaly detection took less than 10 min, with the Gaussian process regression being the rate-limiting step. Even when considering the additional step of producing light curves from images in the first place, the overall run time is low enough to allow for a low-latency online transient detection system. Commensal real-time searches already exist on some interferometers, such as the *realfast* system on the Karl G Jansky's Very Large Array (Law et al. 2018) and on the Murchison Widefield Array (see methods in Hurley-Walker et al. 2023). This work shows that the incorporation of anomaly detection techniques would be a flexible addition to such a backend system, whereby the anomaly ranking and user-defined preference would allow for rapid discovery without ballooning data rates.

However, additional steps will be required in order to maximize potential discoveries from current and upcoming data streams. For example, several recent discoveries have been made using fast images (that is, sub-integrations of a single long observation) that warrant further investigation (Caleb et al. 2022; Hurley-Walker et al. 2022; Wang et al. 2023), while in this work the images are typically separated by a week. We plan to apply these anomaly detection methods to sets of 8s-images taken by MeerKAT in the near future. Similarly, we have not made use of any spectral or polarimetric information inherent in the data, both of which can provide crucial additional information and reduce contaminants (Heywood 2023). However, it is impossible to know on which time-scales and at which frequencies to search in a manner that is completely agnostic to all the possible underlying astrophysics. Therefore it is worth thinking about the specific science case in mind and how to best leverage these data products – for example, if one is searching for radio flaring stars, the use of Stokes V imaging is a powerful tool that eliminates many contaminants from images (e.g. Pritchard et al. 2021, 2024). These additional pipeline steps would, however, add to the computational time of finding these anomalies. It is therefore down to the given science case whether data rates are such that only the bare minimum information can be extracted, or whether the astrophysical phenomena necessitate additional processing.

6.3 Citizen science

All of the discoveries used as our ground truth were found by volunteers. With upcoming (big) data projects such as the SKA and LSST, it is worth discussing the role of individuals and groups with regards to discovery science. A crude but immediate

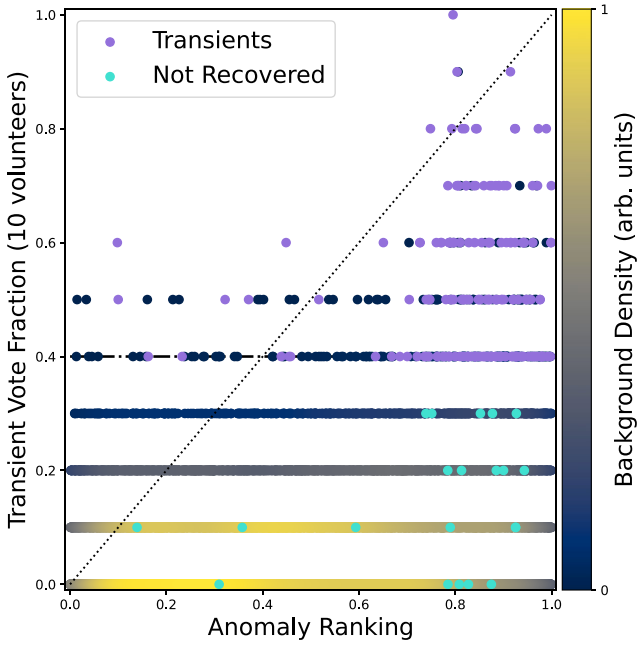


Figure 9. A comparison between how many citizen scientists voted for a source as a transient and its ranking in the anomaly detection work, where a rank of 1 indicates the most anomalous source, whilst the least anomalous source has a ranking of 0. Sources 'not recovered' refers to known variables/transients in the light curves that fall below the 4/10 threshold of A23 and were therefore regarded as missed by volunteers. The background points are all remaining sources, coloured by the density of points. The anomaly list used here corresponds to the wavelet features applied to the LOF algorithm, but results are consistent for all feature-model pairs.

comparison between human discoveries and automated anomaly detection algorithms can be seen in Fig. 9, where the transient vote fraction is how many out of 10 volunteers that saw each source voted for it as a transient/variable. The anomaly ranking is a source's position in the anomaly score-ordered list for the wavelet-LOF instance, though other feature-model pairs were tested with qualitatively similar results. We can see that the majority of volunteer transients are ranked highly by the anomaly detection model, matching the recall curves seen earlier. However, we can also see that there are known transients and variables missed by volunteers (i.e below A23's threshold of 0.4) – these are typically low significance sources and/or those that show variability that was difficult for volunteers to identify but were found in previous studies i.e. Driessen et al. (2020, 2022b) and Rowlinson et al. (2022). It is encouraging then that the anomaly detection methods rank these anomalies as highly as many of the new transients/variables confirmed by volunteers, indicating that we can avoid a potential source of bias in our searches. Similarly, there are sources given low scores by the anomaly detection algorithms (even with active learning) that were clearly identified by a majority of volunteers – these are the data points about the $y = x$ line in Fig. 9. So these are anomalies that volunteers find 'more easily' than our algorithms. It is therefore pertinent to leverage each search method in a complementary manner (see also Wright et al. 2017). For example, one could perform a first pass with anomaly detection procedures, which would find the highly anomalous sources easily and then pass the sources that are not ranked highly (e.g. beyond the 'knee' of a given recall curve) to citizen scientists. However, it is also important to value the time taken by individual vol-

Downloaded from https://academic.oup.com/mnras/article/538/3/1397/8042556 by Universiteit van Amsterdam user on 08 April 2026

unteers and consider whether asking them to find relatively few sources in large data volumes is an optimal or even fair division of labour. We caveat that this brief comparison with volunteer performance may not necessarily generalize to other projects, future data releases or other volunteers and that project builders could in theory give different training and instructions, producing different results. Nevertheless, the combination of rapid computation and the advantages of volunteer-led searches in concert with expert analysis and, as demonstrated in this work, anomaly detection techniques, hold great promise for future transient hunting efforts.

Involving citizen scientists at the active learning stage is one way to leverage this combination of algorithmic efficiency and volunteer diligence. As simulated and then demonstrated by the *Galaxy Zoo* team (Walmsley et al. 2020, 2022a), volunteers can be preferentially shown sources which are most informative for a model to train on, dramatically reducing the number of votes required overall. In an unsupervised setting this human-in-the-loop cycle might constitute showing volunteers the most anomalous sources after default anomaly detection and then updating an acquisition function based on the citizen scientists' classifications. Volunteers could then be shown sources according to the new anomaly score-ordered list and the process repeats. Due to the subjective nature of what a given (citizen) scientist might find interesting, the resulting acquisition function would be some average of many different opinions, weighted by the number of sources classified per volunteer. Project scientists would then have to decide whether or not to intentionally attempt to influence behaviour by, for example, displaying'wanted' source classes in volunteer training, or whether a diversity in voting patterns can be leveraged for downstream tasks. In any case, optimizing what sources are shown to volunteers will be of paramount importance when dealing with future surveys such as from the SKA or Vera C. Rubin Observatory.

7 CONCLUSIONS

In this paper, we have presented the first analysis of anomaly detection models and active learning for finding radio transients via their light curves. We are able to make use of volunteer contributions on the *Bursts from Space: MeerKAT* citizen science project as a ground truth sample against which to test these unsupervised machine learning techniques incorporated into the ASTRONOMALY package. We have shown how choices in feature sets result in different clustering of sources and how the appropriate representation of data has a strong impact on our ability to recover anomalies. We have also shown that choice of anomaly detection model is a less significant determinant in performance. For any combination of feature set and model, the hit rate of transients is always at least a factor of 2 greater than previous methods, for highly anomalous sources. Active learning on a small subset of the data produces better recall of transients, particularly for those ranked as highly anomalous. This enables users to focus on anomalies that are interesting for their given science case. These promising methods, in combination with the volunteer-led efforts and commensal search engines being built into telescope backends, will allow for the unprecedented discovery of rare and novel transients that will drive science cases for upcoming facilities such as the SKA.

ACKNOWLEDGEMENTS

AA acknowledges the support given by the Science and Technology Facilities Council through a STFC studentship and the support of the Breakthrough Listen project. Breakthrough Listen is managed by the Breakthrough Initiatives, sponsored by the Breakthrough Prize Foundation. CJL acknowledges support from the Alfred P. Sloan Foundation. ML acknowledges support from the South African Radio Astronomy Observatory and the National Research Foundation (NRF) towards this research. Opinions expressed and conclusions arrived at, are those of the authors and are not necessarily to be attributed to the NRF. JvdE acknowledges a Warwick Astrophysics prize post-doctoral fellowship made possible thanks to a generous philanthropic donation, and was supported by funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101148693 (MeerSHOCKS) for part of this work. MV acknowledges financial support from the Inter-University Institute for Data Intensive Astronomy (IDIA), a partnership of the University of Cape Town, the University of Pretoria and the University of the Western Cape, and from the South African Department of Science and Innovation's National Research Foundation under the ISARP RADIOMAP Joint Research Scheme (DSI-NRF Grant Number 150551) and the CPRR HIPPO Project (DSI-NRF Grant Number SRUG22031677).

This publication uses data generated via the Zooniverse.org platform, development of which is funded by generous support, including a Global Impact Award from Google, and by a grant from the Alfred P. Sloan Foundation.

MeerKAT is operated by the South African Radio Astronomy Observatory (SARAO), which is a facility of the National Research Foundation, an agency of the Department of Science and Innovation. We thank the SARAO staff involved in obtaining the MeerKAT observations.

We acknowledge the use of the ilifu cloud computing facility – www.ilifu.ac.za, a partnership between the University of Cape Town, the University of the Western Cape, Stellenbosch University, Sol Plaatje University, and the Cape Peninsula University of Technology. The ilifu facility is supported by contributions from the Inter-University Institute for Data Intensive Astronomy (IDIA – a partnership between the University of Cape Town, the University of Pretoria, and the University of the Western Cape), the Computational Biology division at UCT and the Data Intensive Research Initiative of South Africa (DIRISA). This work made use of the CARTA (Cube Analysis and Rendering Tool for Astronomy) software (DOI 10.5281/zenodo.3377984 – https://cartavis.github.io).

This research has made use of the Python programming language and the following open source packages: ASTROPY (The Astropy Collaboration 2013, 2018, 2022), MATPLOTLIB (Hunter 2007), NUMPY (Harris et al. 2020), PANDAS (McKinney 2010), PYWAVELETS (Lee et al. 2019), SCIKIT-LEARN (Pedregosa et al. 2011), SCIPY (Virtanen et al. 2020), and UMAP-LEARN (McInnes et al. 2018; Sainburg, McInnes & Gentner 2021).

DATA AVAILABILITY

ThunderKAT raw data are available on the SARAO archive (https://archive.sarao.ac.za/). Light curves of all data used herein can be found at https://github.com/AnderssonAstro/BfS-MKT-Analysis with digital object identifier 10.5281/zenodo.10462687.

All codebases used for feature extraction, anomaly detection and active learning are open source and can be found from Cabral et al. (2018a, b), Alves et al. (2022), Lochner & Bassett (2021), and Pedregosa et al. (2011) as appropriate.

Downloaded from https://academic.oup.com/mnras/article/538/3/1397/8042556 by Universiteit van Amsterdam user on 08 April 2026

REFERENCES

Alves C. S., Peiris H. V., Lochner M., McEwen J. D., Allam T., Biswas R., 2022, ApJS, 258, 23
Andersson A. et al., 2022, MNRAS, 513, 3482
Andersson A. et al., 2023, MNRAS, 523, 2219
Arfaoui S., Mabrouk A., Cattani C., 2021, Wavelet Analysis: Basic Concepts and Applications. Chapman and Hall / CRC Press, New Yorkhttps://books.google.co.uk/books?id=ZT4oEAAAQBAJ
Braun R., Bonaldi A., Bourke T., Keane E., Wagg J., 2019, preprint (arXiv:1912.12699)
Breiman L., 2001, Mach. Learn., 45, 5
Breuniq M. M., Kriegel H. P., Ng R. T., Sander J., 2000, ACM SIGMOD Record, 29, 93
Bright J. S. et al., 2020, Nat. Astron., 4, 697
Byrd R. H., Lu P., Nocedal J., Zhu C., 1995, SIAM J. Sci. Comput., 16, 1190
Cabral J. B., Sanchez B., Ramos F., Gurovich S., Granitto P., VanderPlas J., 2018a, Astrophysics Source Code Library, record ascl: 1806.001
Cabral J. B., Sánchez B., Ramos F., Gurovich S., Granitto P. M., Vanderplas J., 2018b, Astron. Comput., 25, 213
Caleb M. et al., 2022, Nat. Astron., 6, 828
Caleb M. et al., 2024, Nat. Astron., 8, 1159
Chastain S. I., van der Horst A. J., Rowlinson A., Rhodes L., Andersson A., Diretse R., Fender R. P., Woudt P. A., 2023, MNRAS, 526, 1888
Chastain S. I. et al., 2024, preprint (arXiv:2412.02832)
Czech D., Mishra A., Inggs M., 2018, Astron. Comput., 25, 52
de Ruiter I. et al., 2024, preprint (arXiv:2408.11536)
Debosscher J., Sarro L. M., Aerts C., Cuypers J., Vandenbussche B., Garrido R., Solano E., 2007, A&A, 475, 1159
Driessen L. N. et al., 2020, MNRAS, 491, 560
Driessen L. N., Williams D. R., McDonald I., Stappers B. W., Buckley D. A., Fender R. P., Woudt P. A., 2022a, MNRAS, 510, 1083
Driessen L. N. et al., 2022b, MNRAS, 512, 5037
Driessen L. N. et al., 2024, MNRAS, 527, 3659
Etsebeth V., Lochner M., Walmsley M., Grespan M., 2024, MNRAS, 529, 732
Fender R. et al., 2016, preprint (arXiv:1711.04132), http://pos.sissa.it/
Fijma S. et al., 2024, MNRAS, 528, 6985
Giles D., Walkowicz L., 2019, MNRAS, 484, 834
Harris C. R. et al., 2020, Nature, 585, 357
Heywood I., 2020, Astrophysics Source Code Library, record ascl:2009.003
Heywood I., 2023, MNRAS, 525, L76
Heywood I. et al., 2022, MNRAS, 509, 2150
Hotan A. W. et al., 2021, PASA, 38, e009
Hughes P. A., Aller H. D., Aller M. F., 1992, ApJ, 396, 469
Hunter J. D., 2007, Comput. Sci. Eng., 9, 90
Hurley-Walker N. et al., 2022, Nature, 601, 526
Hurley-Walker N. et al., 2023, Nature, 619, 487
Hurley-Walker N. et al., 2024, ApJ, 976, L21
Ishida E. E. et al., 2021, A&A, 650, A195
Ivezić Ž. et al., 2019, ApJ, 873, 111
Jankowski F. et al., 2022, in Ruiz J. E., Pierfedereci F., Teuben P.eds, ASP Conf. Ser. Vol. 532, Astronomical Data Analysis Software and Systems XXX. Astron. Soc. Pac., San Francisco, p. 273
Jonas J. L., MeerKAT Team T., 2016, in Proceedings of MeerKAT Science: on the Pathway to the SKA. Sissa Medialab Srl, Stellenbosch, p. 1, https://ui.adsabs.harvard.edu/abs/2016mks..confE...1J/abstract
Kim D.-W., Protopapas P., Alcock C., Byun Y.-I., Bianco F. B., 2009, MNRAS, 397, 558
Kim D.-W. et al., 2011, ApJ, 735, 68
Kim D.-W., Protopapas P., Bailer-Jones C. A. L., Byun Y.-I., Chang S.-W., Marquette J.-B., Shin M.-S., 2014, A&A, 566, A43
Law C. J. et al., 2018, ApJS, 236, 8
Lee G. R., Gommers R., Waselewski F., Wohlfahrt K., OLeary A., 2019, J. Open Source Softw., 4, 1237
Liu F. T., Ting K. M., Zhou Z. H., 2008, IEEE International Conference on Data Mining. Pisa, Italy, p. 413
Lochner M., Bassett B. A., 2021, Astron. Comput., 36, 100481
Lochner M., Rudnick L., 2025, AJ, 169, 121
Lochner M., McEwen J. D., Peiris H. V., Lahav O., Winter M. K., 2016, ApJS, 225, 31
Lochner M., Rudnick L., Heywood I., Knowles K., Shabala S. S., 2023, MNRAS, 000, 1
Lomb N. R., 1976, Ap&SS, 39, 447
Malanchev K. L. et al., 2021, MNRAS, 502, 5147
Malenta M. et al., 2020, in Pizzo R., Deul E. R., Mol J. D., de Plaa J., Verkouter H.eds, ASP Conf. Ser. Vol. 527, Astronomical Data Analysis Software and Systems XXIX. Astron. Soc. Pac., San Francisco, p. 457
McInnes L., Healy J., Melville J., 2018, preprint (arXiv:1802.03426), https://arxiv.org/abs/1802.03426
McKinney W., 2010, in Stéfan van der W., Jarrod M., eds, Proceedings of the 9th Python in Science Conference. Austin, Texas, p. 56
Mesarcik M., Boonstra A. J., Iacobelli M., Ranguelova E., de Laat C. T. A. M., van Nieuwpoort R. V., 2023, A&A, 680, A74
Mohale K., Lochner M., 2024, MNRAS, 530, 1274
Mooley K. P. et al., 2016, ApJ, 818, 105
Murphy T. et al., 2013, PASA, 30, e006
Murphy T. et al., 2017, MNRAS, 466, 1944
Muthukrishna D., Mandel K. S., Lochner M., Webb S., Narayan G., 2022, MNRAS, 517, 393
Nun I. et al., 2015, preprint (arXiv:1506.00010)
Pearson K., 1901, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, 2, 559
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Pelisoli I. et al., 2023, Nat. Astron., 7, 931
Peters N., 2023, The First Transient Search with Automatically Created MeerKAT Images. MSc Thesis, University of Amsterdam, Amsterdamhttps://scripties.uba.uva.nl/search?id=record_54068
Pritchard J. et al., 2021, MNRAS, 502, 5438
Pritchard J., Murphy T., Heald G., Wheatland M. S., Kaplan D. L., Lenc E., O'Brien A., Wang Z., 2024, MNRAS, 529, 1258
Protopapas P., Huijse P., Estévez P. A., Zegers P., Príncipe J. C., Marquette J.-B., 2015, ApJS, 216, 25
Pruzhinskaya M. V., Malanchev K. L., Kornilov M. V., Ishida E. E., Mondon F., Volnova A. A., Korolev V. S., 2019, MNRAS, 489, 3591
Pruzhinskaya M. V. et al., 2023, A&A, 672, A111
Rasmussen C. E., Williams C. K. I., 2006, in Dietterich Thomas, ed., Gaussian Processes for Machine Learning. The MIT Press, Cambridge
Rhodes L., Caleb M., Stappers B. W., Andersson A., Bezuidenhout M. C., Driessen L. N., Heywood I., 2023, MNRAS, 525, 3626
Richards J. W. et al., 2011, ApJ, 733, 10
Rogers B., Lintott C. J., Croft S., Schwamb M. E., Davenport J. R. A., 2024, AJ, 167, 118
Rowlinson A. et al., 2019, Astron. Comput., 27, 111
Rowlinson A. et al., 2022, MNRAS, 517, 2894
Sainburg T., McInnes L., Gentner T. Q., 2021, Neural Comput., 33, 2881
Sarbadhicary S. K. et al., 2021, ApJ, 923, 31
Scargle J. D., 1982, ApJ, 263, 835
Smirnov O. M. et al., 2024, MNRAS, 528, 6517
Smirnov O. M. et al., 2025, MNRAS, 538, L89
Sooknunan K. et al., 2021, MNRAS, 502, 206
Stappers B., 2016, in MeerKAT Science: On the Pathway to the SKA. Proceedings of Science, p. 10
Swinbank J. D. et al., 2015, Astron. Comput., 11, 25
The Astropy Collaboration, 2013, A&A, 558, A33
The Astropy Collaboration, 2018, AJ, 156, 123
The Astropy Collaboration, 2022, ApJ, 935, 167
Villar V. A., Cranmer M., Berger E., Contardo G., Ho S., Hosseinzadeh G., Lin J. Y.-Y., 2021, ApJS, 255, 24
Virtanen P. et al., 2020, Nat. Methods, 17, 261
Walmsley M. et al., 2020, MNRAS, 491, 1554

Downloaded from https://academic.oup.com/mnras/article/538/3/1397/8042556 by Universiteit van Amsterdam user on 08 April 2026

Walmsley M. et al., 2022a, MNRAS, 509, 3966
Walmsley M. et al., 2022b, MNRAS, 513, 1581
Wang Y., Tuntsov A., Murphy T., Lenc E., Walker M., Bannister K., Kaplan D. L., Mahony E. K., 2021a, MNRAS, 502, 3294
Wang Z. et al., 2021b, ApJ, 920, 45
Wang Z. et al., 2022, MNRAS, 516, 5972
Wang Y. et al., 2023, MNRAS, 523, 5661
Webb S. et al., 2020, MNRAS, 498, 3077
Webb S. et al., 2021, MNRAS, 506, 2089
Wright D. E. et al., 2017, MNRAS, 472, 1315

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

Downloaded from https://academic.oup.com/mnras/article/538/3/1397/8042556 by Universiteit van Amsterdam user on 08 April 2026

© 2025 The Author(s).
Published by Oxford University Press on behalf of Royal Astronomical Society. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.