Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **543,** 351–357 (2025) https://doi.org/10.1093/mnras/staf1477
Advance Access publication 2025 September 8

Anomaly detection to identify transients in LSST time series data

Miguel Crispim Romão,[1]★ Djuna Croon [1]★ and Daniel Godines[2]★

[1]*Department of Physics, Institute for Particle Physics Phenomenology, Durham University, Durham DH1 3LE, UK*
[2]*New Mexico State University, 1780 E University Ave, Las Cruces, NM 88003, USA*

Accepted 2025 August 29. Received 2025 August 29; in original form 2025 April 4

ABSTRACT

We introduce a novel approach to detecting microlensing events and other transients in light curves, utilizing the isolation forest (IFOREST) algorithm for anomaly detection. Focusing on the Legacy Survey of Space and Time (LSST) by the Vera C. Rubin Observatory, we show that an IFOREST trained on *signal-less* light curves can efficiently identify microlensing events by different types of dark objects and binaries, as well as variable stars. We further show that the IFOREST has real-time applicability through a drip-feed analysis, demonstrating its potential as a valuable tool for LSST alert brokers to efficiently prioritize and classify transient candidates for follow-up observations.

Key words: gravitational lensing: micro – surveys – software: machine learning – stars: variables: RR Lyrae.

1 INTRODUCTION

The upcoming Legacy Survey of Space and Time (LSST) by the Vera C. Rubin Observatory will redefine time-domain astrophysics. LSST's wide-field, deep-imaging strategy will capture an unprecedented volume of light curve data, revealing a diverse array of transient phenomena. Among these are gravitational microlensing events, including by potential dark matter (DM) candidates (Drlica-Wagner et al. 2019), which are characterized by the transient brightening of a background star due to the gravitational field of an intervening object. A central challenge in leveraging LSST's data set is the early recognition of such transient signals, which is critical for initiating timely follow-up observations.

Recent advancements in outlier and anomaly detection for transient surveys have leveraged machine learning techniques to identify rare or novel astrophysical events in real-time (e.g. Muthukrishna et al. 2022; Aleo et al. 2024; Gupta, Muthukrishna & Lochner 2024). In preparation for LSST, efforts have been made to develop classification algorithms tailored to its alert stream (Soraisam et al. 2020), as well as deep-learning models capable of identifying transients in real time (Shah et al. 2025).

In previous work (Godines et al. 2019), one of us introduced the use of a machine learning classifier to search for microlensing in wide-field surveys with low cadence data. In a recent work (Crispim Romão & Croon 2024), two of us adapted this technique to search for different dark objects, focusing on two representative classes: boson stars, with flat density profiles leading to characteristic light curves with caustic peaks, and Navarro–Frenk–White (NFW) subhalo profiles, which are more sharply peaked and do not exhibit such features in their light curves.

In this work, we pioneer the use of anomaly detection by means of an isolation forest (IFOREST) to identify microlensing events and other transients in LSST light curves. We utilize the cadence simulations provided by the Rubin Observatory to accurately reproduce the transient signals that will be observed during the LSST. These Operational Simulations ('OpSims', Bianco et al. 2021)[1] provide simulated observations using the Rubin scheduler, thus allowing the community to assess different survey strategies so as to optimize survey parameters for maximum scientific output, including the cadence and sky coverage of the survey. The defining innovation in this work is that we also use 'OpSims' to simulate a *signal-less* or 'constant' class, which we use to train our IFOREST model on. We find that an IFOREST trained on signal-less light curves can effectively identify lensing events due to point-like as well as extended lenses, as well as binary lenses and variable stars such as RR Lyrae.

Importantly, we demonstrate that this technique has potential for an early detection system that could predict transient events before they reach their peak brightness. Early prediction is critical for timely follow-up observations and for maximizing the scientific yield of detected events. We propose a framework for such a system, leveraging early identification to optimize resource allocation for follow-up studies.

This paper is organized as follows. In Section 2, we discuss the sensitivity of LSST to microlensing phenomena, we introduce the classes of transients that we will be discussing throughout this paper, and we present the details of the simulated data generated for our analysis. In Section 3, we discuss the detection prospects of microlensing events using machine learning. This study is broken down into three analyses: the first uses an anomaly detection model to identify promising light curves and is presented in Section 3.1; next, in Section 3.2, we perform a drip-feed analysis to assess how the

★ E-mail: miguel.romao@durham.ac.uk (MR); djuna.l.croon@durham.ac.uk (DC); godines@nmsu.edu (DG);

[1]https://www.lsst.org/content/charge-survey-cadence-optimisation-committee-scoc

anomaly detection model could be used as an online alarm to help brokers identify potential microlensing events; the final analysis, presented in Section 3.3, employs a classifier to further help identify promising light curves that have passed the anomaly detection alarm, this will allow us to discuss the expected LSST sensitivity to different classes of transients. In Section 4, we conclude and discuss the prospects of the potential detection of extended objects by the LSST.

2 LIGHT CURVES

2.1 Light curve classes

We generate light curves for five objects, including transients and RR Lyrae variables stars. In addition, we simulate a constant class (Constant) to mimic real data from non-variable sources. This class represents signal-less observations and is used to train the anomaly detection model. Due to the complexity of incorporating color information, we restrict our analysis to single-band light curves, especially the *i* band.

The lensing classes are comprised of light curves from point-like microlensing events (ML), binary microlensing (Binary ML), and microlensing by extended objects: NFW subhaloes and boson stars (BS). These classes represent a diverse collection suitable for identifying various DM objects, including primordial black holes (PBHs), extended objects with sharply peaked density profiles (NFW subhaloes), and those with flat density profiles (BS) set against a realistic astrophysical background. In this study, we neglect microlensing parallax, which is expected to be small for short duration events. In Abrams et al. (2025), it was found that parallax parameter π_E was detected with 2σ confidence level for a small fraction of events for $t_{\rm E} \sim 100$, on the edge of our simulation range. We note that parallax can provide information about the proper motion and distance of the lens, allowing for a more precise measurement of the mass of the lens which may be explored in future work.

2.2 Data set generation

For the Constant, variable, and point-like microlensing events classes, we utilize the simulation module provided in MICROLIA (Godines et al. 2019), while binary lenses are simulated using PYLIMA, an open-source software for microlensing analysis and simulation Bachelet et al. (2017). MICROLIA simulates variable stars using GATSPY's template-based fitting method to model variable stars using real RRLyrae templates (VanderPlas & Ivezić 2015; VanderPlas 2016). For the extended lenses, NFW and BS, we compute the magnification using a method introduced in Croon, McKeen & Raj (2020a) and Croon et al. (2020b) with the code used in Crispim Romão & Croon (2024). Single-lens events are characterized by three key parameters that describe the lens-source approach: the Einstein crossing time ($t_{\rm E}$), the time of maximum magnification (t_0), and the minimum impact parameter (u_0), which quantifies the separation between source and lens at $t = t_0$. The time t_0 is randomly drawn between the 1st and 99th percentiles of the observation window, extended by half the Einstein crossing time on either side to allow for values slightly outside the observed data range. However, this may result in signal-less simulations if t_0 falls within a period when the point in the sky was not visited. We emphasize that our sampling strategy for the microlensing parameters was deliberately chosen to serve the objectives of this work. A fully self-consistent Galactic model would introduce the source-lens relative proper motion $\mu_{\rm rel}$ through $t_E = \theta_E/\mu_{\rm rel}$, allowing an astrophysically motivated sampling over the distributions of lens masses, distances, and velocities. Our aim here, however, is not to reproduce a realistic population, but rather to generate a broad and controlled coverage of possible light curve shapes for training and testing classification algorithms. For this purpose, sampling directly in (t_E, u_0, t_0) space is efficient and sufficient.

Table 1. Parameters used in the simulation of the lensing classes: Einstein crossing time $t_{\rm E}$, minimal impact parameter (normalized to the Einstein radius) u_0, peak time t_0, source star flux $F_{\rm s}$.

parameter	min	max	spacing
t_E (d)	0[a]	100	linear
u_0	0	3	linear
t_0	$t_{1\ \text{per cent}} - 0.5t_E$	$t_{99\ \text{per cent}} + 0.5t_E$	linear
ρ[b]	0	0.05	log
s[b]	0.3	3	log
q[b]	0	1	linear
α (rad)[b]	0	2π	linear
τ_m[c]	0.05	5	log
F_s	0	1	linear

Notes. [a]While $t_{\rm E} = 0$ d is physically meaningless, the probability of drawing this exactly is vanishingly small and such events do not appear in our data set.
[b]These parameters define the binary-lens system: the angular radius of the source star ρ, the projected separation between the two lens masses s, the mass ratio of the binary lenses q, the angle between the source's trajectory and the binary lens axis α.
[c]Only used to simulate the extended dark objects: normalized dark object size $\tau_{\rm m}$.

The blending fraction, $F_{\rm s}$, representing the fraction of the object's flux affected by lensing, is given by

$$F_{\rm s} = \frac{F_{\rm source}}{F_{\rm baseline}}, \tag{1}$$

and is randomly between 0 and 1, where $F_{\rm s} = 0$ corresponds to a scenario where the source is significantly blended.

In the case of static binary lenses where a secondary body is present, the light curve can be characterized by four additional parameters. These parameters include lens separation s, mass ratio q, source size ρ, and source trajectory angle α, which determines how the source crosses the caustic structures (for a detailed discussion; see Dominik 1998). Our choice of a linear spacing in q is motivated by the goals of this initial study: rather than reproducing an astrophysically realistic binary population, we seek to map out the diversity of microlensing morphologies. In particular, near $q \to 1$ the light curve deviates strongly from the point-lens form and produces symmetric caustic structures, resembling some of the extended DM candidates we consider. A linear sampling ensures sufficient coverage of this regime, enabling us to test how well such cases can be distinguished from exotic alternatives. For the BS and NFW subhaloes, we adopt the light curves described in Crispim Romão & Croon (2024), using the same ranges for the single-lens parameters listed in Table 1. Additionally, we include the lens size normalized to the Einstein radius, defined as $\tau_{\rm m} \equiv r_{\rm lens}/r_{\rm E}$, which is sampled over the range $0.05 < \tau_{\rm m} < 5$.

The above parameters set the theoretical magnification of the light curve brightness by different dark objects and variable sources. In order to study the prospect of detection of dark objects by the LSST, we need to generate realistic light curves. To do so, we need to simulate the survey cadence and data acquisition pipeline. To this effect, we simulated the light curves using the `baseline_v2.0_10yrs` Baseline Survey Strategy, published as part of 'OpSims' and included in `rubin-sim`. We extract the

5σ depth photometry for the 'i' band using 30-s exposures, and a zeropoint of 27.85. The observation times are position-dependent as the observation strategy is inhomogeneous across the LSST footprint. As we seek to explore classification performance with sparsely sampled photometry, we select the cadences for each simulated light curve via a random selection of sky positions. This ensures uniform coverage across the entire survey, with the bulk of the observations coming from the Deep-Wide-Fast (WFD) survey mode which will comprise 90 per cent of the LSST observing time.

The data set was split into train, validation, and test subsets with relative proportions 0.5:0.25:0.25. The training set was used for exploratory data analysis and to train the machine learning models. The validation set was used for hyperparameter optimization (when applicable) and to produce a preliminary analysis. The test set was only used to produce the final analyses presented in the next section. The data set and the artefacts for the trained models are available at Crispim Romão, Croon & Godines (2025)[2].

3 DETECTION PROSPECTS

We now study the detection prospect for the classes of light curves presented in the preceding section. First, in Section 3.1, we demonstrate how an anomaly detection model can be used to filter out signal-less light curves, providing a purely data-driven approach to candidate selection. Then, in Section 3.2, we develop a drip-feed analysis similar to that in Godines et al. (2019) to assess how this filtering strategy can enable the early identification of transient events in the LSST data stream, potentially triggering follow-up observations of the source. Finally, in Section 3.3, we conduct a classification analysis, following the approach of Crispim Romão & Croon (2024), to investigate what can be inferred about these events based solely on LSST data, regardless of whether follow-up observations have been performed.

3.1 Anomaly detection

The first step of our analysis aims to develop a data-driven methodology that can filter out signal-less light curves to identify the most promising candidate light curves for further study. To this effect, we will train an isolation Forest (IFOREST) Liu, Ting & Zhou (2008) on signal-less light curves – more precisely, those from the Constant class described in the previous section. This approach enables the IFOREST to identify anomalies as light curves more likely to contain a signal, either from variable sources or lensing phenomena.

An IFOREST is a collection of N trees $\{T_i\}$, where $i = 1, \ldots, N$, grown through random recursive partitions of a data sub-sample. The trees will grow until they reach a maximum depth or can no longer partition further (i.e. all leaf nodes have a single data point). Because the partitions are random, the number of nodes that a data point traverses from root to leaf nodes is an indication of how inline it is, with anomalies being identified as 'outliers' which are more likely to arrive at a leaf node on a shorter path than an 'inlier', i.e a signal-less light curve. Once the forest is grown, a score can be computed on a data point, x, as

$$iF\text{score}(x) = 0.5 - 2^{-\frac{\mathbb{E}[h(x)]}{c}}, \quad (2)$$

where $\mathbb{E}[h(x)] = \frac{1}{N}\sum_i h_i(x)$ is the average depth at which x is found across all tress in the forest, with $h_i(x)$ denoting the depth of x in tree T_i. The constant c represents the average depth of a data point in a binary tree with the same maximal depth as the trees $\{T_i\}$. It is clear that $-0.5 \leq iF\text{score}(x) \leq 0.5$, with

(i) $\mathbb{E}[h(x)] \to 0 \Rightarrow iFscore(x) \to -0.5$, i.e. 'outliers' (the anomalies) score negative values as they traverse fewer nodes than an 'inlier'. Data completely out-of-distribution will be immediately isolated and will score near -0.5,

(ii) $\mathbb{E}[h(x)] \sim c \Rightarrow iFscore(x) \sim 0$, i.e. 0 happens when the number of crossings is around the expected average and therefore is a reasonable cut-off between 'inliers' and 'outliers',[3]

(iii) $\mathbb{E}[h(x)] \to \inf \Rightarrow iFscore(x) \to 0.5$, i.e. 'inliers' score positive values as they traverse more nodes than an 'outlier'. Data very close to the centre of mass of the multivariate distribution will traverse the most nodes and will score near 0.5. Notice that the trees are grown to a finite maximum depth; therefore, $\mathbb{E}[h(x)]$ is in practice finite and therefore the 'inlier' scores are expected to be positive but closer to 0.

In this work, the IFOREST will be trained solely on the signal-less Constant class. The IFOREST then acts as a *one-class classifier*, identifying non-constant light curves, i.e. those exhibiting any signal sufficiently distinct from signal-less Constant ones, as anomalies. This methodology is potentially sensitive to *any* type of physical phenomena, not just to candidate signal classes introduced earlier in the previous section, which will be further studied in Section 3.3[4]. A similar methodology using IFOREST has been explored in the context of model agnostic New Physics searches in collider experiments (Crispim Romão, Castro & Pedro 2021). We use SCIKIT-LEARN (Pedregosa et al. 2011) implementation of the IFOREST, and we will leave all hyperparameters set to default[5].

In Fig. 2, we present the histogram of the iForest anomaly scores predictions for different lensing classes (ML, NFW, Binary, and BS), RRLyrae variables, and Constant (signal-less) light curves. Recalling that the simulated light curves of lensing classes might not feature a signal if the event took place outside of a visiting window, we can observe that IFOREST output distributions for constant and lensing classes reveal a mode of lensing light curves corresponding to cases without detectable signals during data acquisition. Additionally, we see how the values of the IFOREST anomaly scores for the RRLyrae are very negative, making this class especially easy to be captured by the IFOREST filter. This is easy to understand: RRLyrae are variable sources which the magnitude varies for large periods of time, producing light curves distinctively different from signal-less Constant light curves, with only a small fraction not exhibiting a potential signal. From the constant test set, we find that 98 per cent of objects are assigned IFOREST scores ≥ 0. Among BS objects, 45 per cent have scores < 0, while for NFW, ML, and Binary classes,

[2]https://zenodo.org/records/15005108

[3]Notice that this does not mean that half the training set will be on each side of 0. c is not computed over the forest grown using the training data, but rather from an estimate for generic binary trees of the same depth. See Liu et al. (2008) for more details.

[4]In the early stages of our work, we also considered Cepheid variables with periods of $\sim$ 3–60 d, included in earlier studies (Godines et al. 2019; Crispim Romão & Croon 2024). However, the IFOREST model easily isolated them because MICROLIA simulated them using a limited set of templates from real RRLyrae, which inherently have low periods ($\lesssim$ 1 d) and different light curve structure such as sawtooth-like asymmetries (e.g. McWilliam 2011). This highlights how unrepresentative simulations can significantly affect sensitivity analyses, potentially leading to overly optimistic assessments.

[5]Due to the lack of a well-defined metric for hyperparameter optimization in semi-supervised outlier detection, we leave hyperparameters at their default values. The impact of the hyperparameters on the sensitivity to different signals is left for future work.

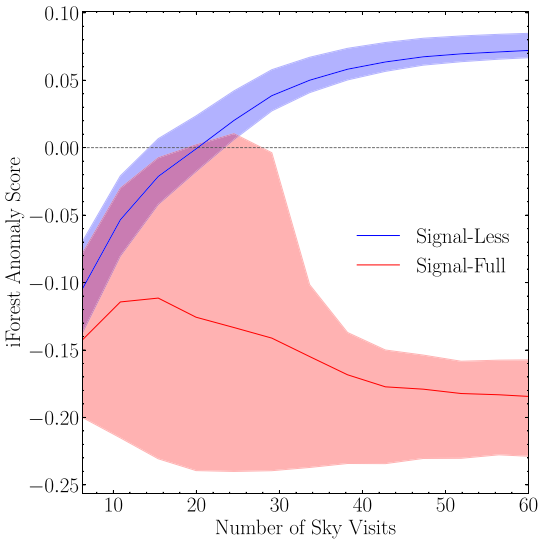

Figure 1. Comparison of anomaly scores for light curves with (this includes microlensing events with several lens profiles (point-like, boson star, NFW subhaloes, and binaries) and without signals. The solid lines show mean bin values, while shaded regions represent the 25th–75th percentile range.

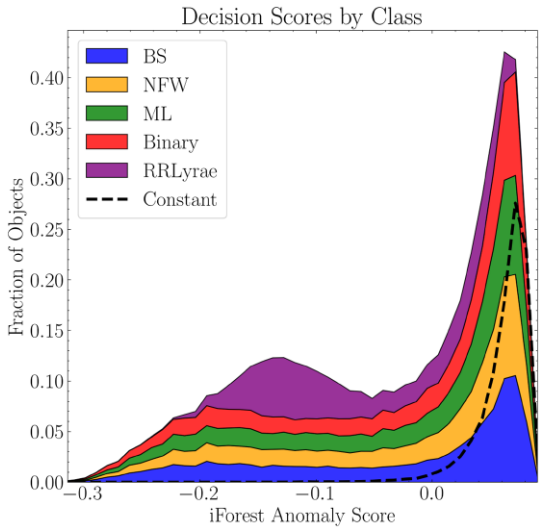

Figure 2. Output of the IFOREST for the different classes.

this fraction is 45 per cent, 46 per cent, and 46 per cent, respectively. For RRLyrae variables, 71 per cent of objects have scores < 0.

These results highlight the IFOREST effectiveness in filtering signal-less light curves in a data-driven methodology. However, it is crucial to examine the model sensitivity to different astrophysical lensing parameters. In Fig. 3, we show the impact of the minimal impact parameter, u_0, the lensing event time, $t_{\rm E}$, and the baseline magnitude of the source on the prediction of IFOREST, where blue (red) represents negative (positive) IFOREST scores[6]. As expected, the IFOREST is more sensitive to lensing events with small minimal impact parameter, $u_0 \lesssim 1.5$, which produces a higher brightness magnification; longer crossing times, $t_{\rm E} \gtrsim 30$ d, which increase the likelihood of detection during observational visits; and brighter sources, magnitude $\lesssim 30$,

[6]We notice that there is a visual bias towards positive values of the IFOREST anomaly score as these points are drawn last and therefore on top.

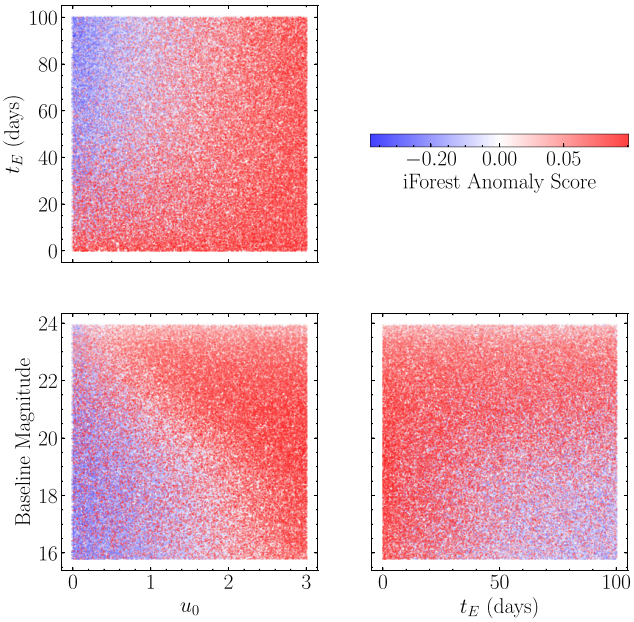

Figure 3. IFOREST Anomaly Score dependency on astrophysical lensing parameters: minimal impact parameter, u_0, Einstein crossing time, t_E, and source baseline magnitude. Blue (red) represents negative (positive) IFOREST anomaly score.

which produces a higher signal to noise ratio. We emphasize that these parameters were not used when training the IFOREST, which was trained only on light curve-derived statistics.

So far, we have explored iForest as a data-driven tool for filtering signal-less light curves, enabling the identification of promising signal-containing candidates with minimal prior assumptions. However, our analysis has been purely offline, relying on full light curves simulated over a two-year span using the `rubin-sim baseline_v2.0_10yrs` survey strategy for LSST. In the next section, we assess IFOREST's feasibility as an online filter for early detection, investigating its ability to identify transient events in real time during data acquisition. Given the results in Fig. 2, we focus on the lensing classes, as RRLyrae light curves tend to cluster around negative IFOREST anomaly scores, making them easier to isolate.

3.2 Online early detection

As is clear from Fig. 3, the IFOREST appears particularly sensitive to lensing phenomena when the minimal impact parameter is low, the Einstein crossing time is long, and the source is bright. To demonstrate its sensitivity to signal-containing light curves, we focus on this region of the lensing parameter space. For this purpose, we define a *sensitivity flag* that selects light curves satisfying the following conditions:

$$\{u_0 < 1 \ \wedge \ t_{\rm E} > 40\,{\rm d} \ \wedge \ \text{Baseline Magnitude} < 19\}\,, \qquad (3)$$

where the values can be adjusted to increase parameter space coverage. This flag serves to illustrate IFOREST potential in detecting interesting light curves, which becomes evident in this region.

In Fig. 4, we show the distribution of IFOREST anomaly scores across the four lensing classes, further subdivided by whether the light curves fall within the region defined by equation (3). The vast majority of lensing light curves in this sensitivity region have negative IFOREST anomaly scores, confirming IFOREST's heightened sensitivity in this subset of the parameter space. However, some light

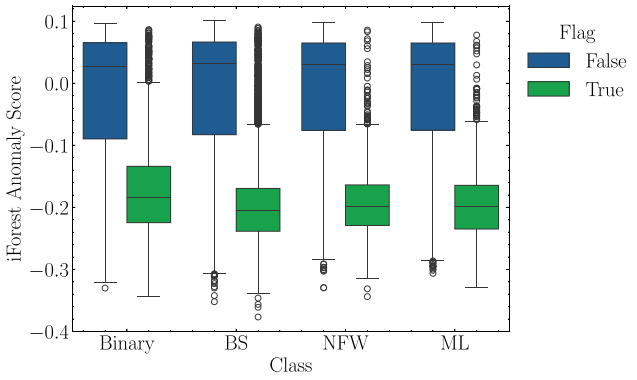

Figure 4. IFOREST Anomaly Score distribution for the four lensing classes, further subdivided on whether they fall in the region equation (3) or not.

curves still yield positive scores, suggesting that even within this region certain events may appear signal-less – likely because their observational coverage does not align with the timing of the transient event and therefore no brightness magnification is observed in the light curve. Additionally, we observe that BS, NFW, and ML light curves have similar distributions, whereas Binary light curves exhibit less negative iForest anomaly scores. The former observation reflects the fact that BS, NFW, and ML light curves, which are single-lens sources, resemble each other more closely than they do Binary light curves, a pattern explored further in Section 3.3. The latter can be attributed to the larger parameter space for Binary events compared to BS, NFW, and ML (see Table 1), which dilutes the number of signal-containing Binary light curves.

To evaluate the early detection capabilities of our classification engine, we conduct a drip-feeding analysis similar to the one presented in Godines et al. (2019), in which the test light curves are classified one timestamp at a time. This approach effectively simulates real-time classifier performance as it ingests new data points and updates IFOREST predictions incrementally. Since too few data points can lead to unstable statistics, we must determine the optimal minimum number of observations (i.e. sky visits) required for reliable online anomaly detection.

In Fig. 1, we illustrate how the IFOREST anomaly score evolves with the number of sky visits. Blue represents signal-less light curves from the Constant class. We observe that when the number of visits is below approximately 20, IFOREST predictions are unreliable, often yielding negative scores that falsely indicate potential signals. However, for 30 or more visits, signal-less Constant light curves are correctly classified, suggesting $\sim$ 20 timestamps are required to confidently filter out signal-less light curves, but $\sim$ 30 to confidently capture transient phenomena. This demonstrates the difficulty in robustly detecting transient phenomena with non-uniform cadence. The seasonal observational gaps, for example, have been shown to significantly reduce the performance of machine learning models when compared to more homogenous year-round cadence (e.g. Fagin et al. 2025). For 30 or more visits, the scores for the microlensing classes stabilize at negative values around -0.15, indicating that these light curves are correctly identified as non-Constant, i.e. as potentially yielding signals. This also happens to the RR Lyrae light curves, which we present in Fig. 5, where we can see that the predictions stabilize for 30 or more sky visits, with only a small portion of RR Lyrae light curves being classified as not having signal in agreement with Fig. 2. We do not consider the IFOREST output to be reliable before stabilization, even though the scores of signal-less and RR Lyrae light curves are separated before this. Note that no

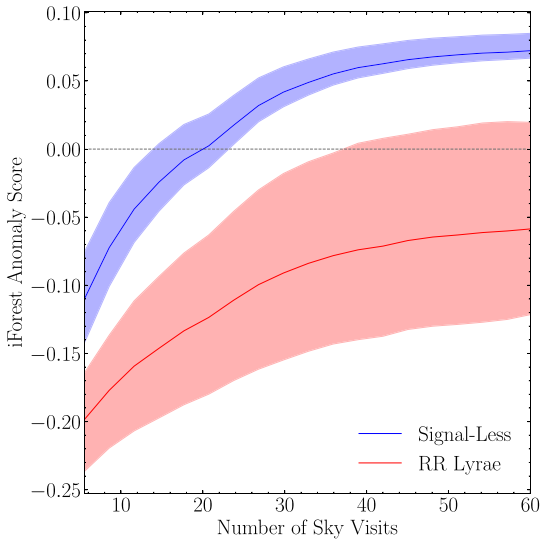

Figure 5. Comparison of anomaly scores for light curves with RR Lyrae and signal-less light curves. The solid lines show mean bin values, and the shaded regions show the 25–75 percentile range.

cuts similar to (3) have been made on the RR Lyrae parameter space, such that the percentile range band is greater in this case.

3.3 Offline signal classification

So far, we have examined how a semi-supervised anomaly detection method based on IFOREST can help *remove* signal-less light curves, allowing us to isolate potential signals. In this section, we investigate the types of signals that could potentially be detected in LSST after applying IFOREST filtering. To this end, we perform an offline classification on light curves with an IFOREST score lower than 0. We exclude Constant class light curves, as the goal of this section is to evaluate the feasibility of classifying various physical phenomena.

We follow a similar analysis to that presented in Crispim Romão & Croon (2024), where a Histogram-based Gradient Boosting Classifier (HGBC) was used. In this study, we focus only on the transient and variable phenomena introduced in Section 2, specifically BS, NFW, point-like ML, Binary ML, and RRLyrae. We employ the SCIKIT-LEARN (Pedregosa et al. 2011) implementation of the HGBC and optimized its hyperparameters using OPTUNA (Akiba et al. 2019). See Section A for more details.

In Fig. 6, we present the confusion matrix obtained by training the HGBC on this data set. A key observation is that microlensing events (BS, NFW, ML) are highly mixed with one another, suggesting that these classes serve as irreducible background contaminants for each other, making their separation challenging. The classifier tends to default to the point-like ML class when struggling to differentiate among these three categories. This is particularly problematic for isolating NFW light curves, as only about 27 per cent are correctly identified, with a significant fraction (33 per cent) being misclassified as ML. On a positive note, BS is apparently more distinguishable than NFW, this might indicate a potential for positive detection, which we explore further below. Additionally, Binary light curves are the easiest transient phenomena to classify, which is expected given their diverse parameter space. This results in a heterogeneous sample with distinct morphologies, making them relatively unique. However, some Binary light curves lack characteristic features and are instead classified as point-like ML. Finally, RRLyrae is the easiest class to

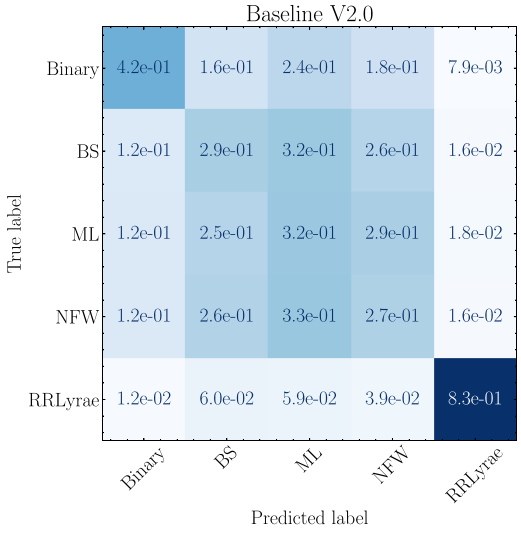

Figure 6. Confusion matrix for the classification task performed on the data set with `rubin-sim` LSST baseline cadence simulation after performing the IFOREST filtering.

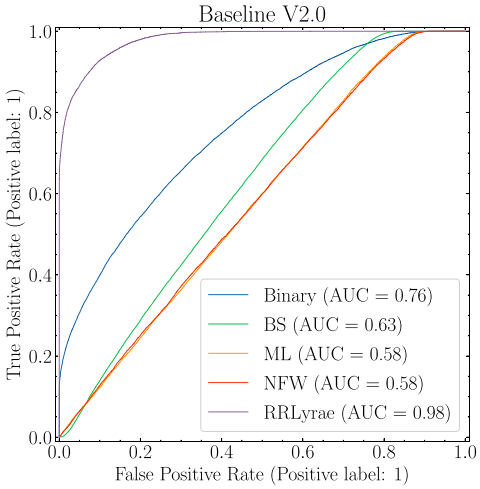

Figure 7. ROC curves and their AUC for all classes for the classification task performed on the data set with `rubin-sim` LSST baseline cadence simulation.

isolate, as its light curves exhibit long-period variations absent in lensing phenomena.

In Fig. 7, we present the Receiver Operator Characteristic (ROC) curves and their corresponding Areas Under the Curve (AUC) values for the trained HGBC. We notice that the classification task is to disentangle different classes, so that the ROC AUC for each class is obtained for that class being assigned the positive label and all other classes to the negative class (one-versus-rest discrimination).

Unfortunately, we observe that for both NFW and BS, there is no regime in which the False Positive Rate vanishes while maintaining a non-zero True Positive Rate. This indicates that it is not possible to unambiguously distinguish NFW and BS light curves from point-like ML using the classifier predictions. This finding contrasts with the results of Crispim Romão & Croon (2024), where the analysis was performed using the OGLE-II survey parameters[7]. We attribute this difference in results to the sparse cadence of the LSST compared to the OGLE survey, which was specifically designed to detect microlensing events.

4 DISCUSSION

This work presents anomaly detection as a means of detecting transient events in time series data. Focusing on the LSST by the Vera C. Rubin Observatory, we have shown that training an IFOREST on signal-less light curves can identify transient events in all categories we simulated. This includes microlensing by DM objects: point-like DM objects, boson stars which have a flat density profile, and NFW subhaloes which have an extended but peaked profile.

Importantly, our methodology can be used in a real time analysis, allowing for the early identification of transient events. This will crucially allow for follow-up strategy of the data produced by a wide-field, low cadence survey like the LSST.

We also tested the potential of the IFOREST to distinguish between different classes of light curve. We found that, unlike previous studies, there is no classification regime where extended lenses can be unambiguously identified using the classifier. However, the higher sensitivity to BS hints at the possibility of a positive detection of these objects in conjugation with a more detailed analysis.

Since microlensing is fundamentally an achromatic phenomenon, our primary focus in this work is on *i*-band observations, which provide a clean and efficient means of detecting and characterizing events without the added complexity of multiband analysis. Color information can provide valuable insights into source properties, blending effects, and potential chromatic signatures in more complex microlensing scenarios, such as binary lenses or finite-source effects. In future work, we will incorporate multiband photometry to explore these additional opportunities. Moreover, while our current simulation neglects microlensing parallax – given its expected minimal impact on the short-duration events analysed – future iterations of our data set will incorporate parallax effects. This will enable us to explore how parallax might help infer lens properties such as proper motion, distance, and mass, particularly in extended or binary lenses.

In this work, we modelled single-lens events using the standard Paczyński point-source model as implemented in MicroLIA, which does not consider finite-source effects. By contrast, our binary-lens modelling with pyLIMA incorporates the source-size parameter ρ which was sampled as per Table 1. We note, however, that ρ can in principle play a role in high-magnification single-lens events (e.g. Yoo et al. 2004), where finite-source effects round off the peak magnification and enable a measurement of the angular Einstein radius. While the LSST cadence and signal-to-noise that characterizes our present data set is largely insufficient to constrain ρ except in high-magnification events observed in well-sampled fields, in future work, we plan to incorporate finite-source effects consistently across both single and binary-lens simulations. This will ensure consistent treatment of microlensing parameters across different event classes.

The anomaly detection methodology developed in this work is based on an IFOREST trained on an offline tabular data set composed of statistical and derived quantities from light curves, and in Section 3.1,

[7]Although Fig. 7 demonstrates that BS light curves are challenging to isolate, the BS ROC curve exhibits a higher AUC than those for the ML and NFW classes, suggesting that the classifier has greater discriminative power to identify BS light curves. As noted in Crispim Romão & Croon (2024), this is due to the emergence of symmetric caustics around the main brightness peak, a distinctive feature of the BS light curves.

we showed that this 'offline' approach is effective at finding signal-full light curves. In Section 3.2 we also evaluated IFOREST in a 'drip-feed' scenario to assess its performance in an online detection setting. In particular, our results indicate that at least 20 timestamps are needed before IFOREST predictions become reliable at rejecting signal-less light curves, highlighting a potential shortcoming in producing a timely alarm for early detection. Future work will explore alternative methods to mitigate this delay by extending our methodology with time-series anomaly detection methods and machine learning techniques designed for online learning over data streams. However, we notice that representing the data in a tabular format remains advantageous, as it condenses key features into a structured form for robust classification, while the sparse and irregular observation cadence of LSST is likely to negatively impact the performance of online methods. A detailed comparison is left for future work.

Finally, this work demonstrates the potential for anomaly detection methodology to detect transient events in the LSST data stream. Therefore, the primary future direction for this research programme is to integrate the techniques in this work directly into LSST science brokers. This can then be applied to scheduled data releases, starting summer 2025, for offline data mining for microlensing and other transient signals.

ACKNOWLEDGEMENTS

We note that authorship ordering on this work is alphabetical; all authors have made important contributions to this work.

We thank Etienne Bachelet for providing guidance on simulating binary microlensing using PYLIMA, and the Rubin Observatory microlensing subgroup for useful discussions. DC thanks the CERN theory group for hospitality during the final stages of this work. MCR and DC are supported by the STFC under Grant No. ST/T001011/1.

DATA AVAILABILITY

The data set used in this work was generated by simulating light curves using publicly available open-source software packages (Bachelet et al. 2017; Godines et al. 2019; Bianco et al. 2021; Crispim Romão & Croon 2024). The data set and the artefacts for the trained models, as well as instructions on how to use them, are available at Crispim Romão et al. (2025).[8]

REFERENCES

Abrams N. S. et al., 2025, ApJS, 276, 23
Akiba T., Sano S., Yanase T., Ohta T., Koyama M., 2019, Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Association for Computing Machinery, New York, USA
Aleo P. D. et al., 2024, ApJ, 974, 172
Bachelet E., Norbury M., Bozza V., Street R., 2017, AJ, preprint (arXiv:1709.08704)
Bianco F. B. et al., 2021, ApJS, 258, 1
Crispim Romão M., Croon D., 2024, Phys. Rev. D, 109, 123004
Crispim Romão M., Castro N. F., Pedro R., 2021, Eur. Phys. J. C, 81, 27
Crispim Romão M., Croon D., Godines D., 2025, LSST light curves for constant and variable sources, and for point-like and extended objects microlensing, Zenodo. Available at: https://doi.org/10.5281/zenodo.15005108
Croon D., McKeen D., Raj N., 2020a, Phys. Rev. D, 101, 083013
Croon D., McKeen D., Raj N., Wang Z., 2020b, Phys. Rev. D, 102, 083021
Dominik M., 1998, A&A, 329, 361
Drlica-Wagner A. et al., 2019, preprint(arXiv e-prints)
Fagin J. et al., 2025, ApJ, 981, 61
Godines D., Bachelet E., Narayan G., Street R., 2019, Astron. Comput., 28, 100298
Gupta R., Muthukrishna D., Lochner M., 2024, preprint (arXiv:2408.08888)
Liu F. T., Ting K. M., Zhou Z.-H., 2008, "Isolation Forest," 2008 Eighth IEEE International Conference on Data Mining. IEEE, Pisa, Italy, p. 413
McWilliam A., 2011, Carnegie Observatories Astrophysics Series, Vol. 5, RR Lyrae Stars, Metal-Poor Stars, and the Galaxy. Carnegie Institution of Washington, Pasadena, CA
Muthukrishna D., Mandel K. S., Lochner M., Webb S., Narayan G., 2022, MNRAS, 517, 393
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Shah V. G., Gagliano A., Malanchev K., Narayan G., 2025, preprint(arXiv e-prints)
Soraisam M. D. et al., 2020, ApJ, 892, 112
VanderPlas J., 2016, gatspy: Version 0.3 Feature Release. Available at: https://zenodo.org/record/47887
VanderPlas J. T., Ivezić Ž., 2015, ApJ, 812, 18
Yoo J. et al., 2004, ApJ, 603, 139

APPENDIX: HYPERPARAMETER OPTIMIZATION

While there is no well defined semi-supervised validation metric for tuning the hyperparameters of IFOREST, we can optimize the multiclass classifier used in the offline analysis to better distinguish light curves from different physical phenomena. We optimized the HGBC hyperparameters using OPTUNA (Akiba et al. 2019), training an HGBC for each proposed hyperparameter combination on the training set and selecting the optimal configuration based on the mean ROC AUC computed on the validation set. The range of hyperparameters explored during optimization, along with their final optimal values, is listed in Table A1. We employed OPTUNA's default optimizer, conducting a maximum of 100 trials.

Table A1. Hyperparameter ranges and optimal values found during the hyperparameter optimization step.

Hyperparameter	Range	Optimal value
`learning_rate`	[0.01, 1]	0.032
`max_iter`	[50, 200]	200
`max_leaf_nodes`	[20, 40]	36
`min_samples_leaf`	[10, 30]	26
`l2_regularisation`	$[10^{-4}, 1]$ or 0	0
`n_iter_no_change`	[1, 20]	18

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

[8] https://zenodo.org/records/15005108