Microlensing signatures of extended dark objects using machine learning

Miguel Crispim Romao[*] and Djuna Croon[†]

*Institute for Particle Physics Phenomenology, Department of Physics, Durham University,
Durham DH1 3LE, United Kingdom*

This paper presents a machine learning-based method for the detection of the unique gravitational microlensing signatures of extended dark objects, such as boson stars, axion miniclusters and subhalos. We adapt MicroLIA, a machine learning-based package tailored to handle the challenges posed by low-cadence data in microlensing surveys. Using realistic observational time stamps, our models are trained on simulated light curves to distinguish between microlensing by pointlike and extended lenses, as well as from other object classes which give a variable magnitude. We focus on boson stars and Navarro-Frenk-White (NFW) subhalos and show that the former, which are examples of objects with a relatively flat mass distribution, can be confidently identified for $0.8 \lesssim r/r_E \lesssim 3$. Intriguingly, we also find that more sharply peaked structures, such as NFW subhalos, can be distinctly recognized from point lenses under regular observation cadence. Our findings significantly advance the potential of microlensing data in uncovering the elusive nature of extended dark objects. The code and dataset used are also provided.

I. INTRODUCTION

Macroscopic dark matter candidates, with masses ranging from large asteroids ($\sim 10^{-15} M_\odot$) to stars ($\sim M_\odot$), offer compelling alternatives to the traditional particle-based theories. These celestial objects, potentially formed in the early Universe, are primarily detectable via their gravitational effects: gravitational lensing (e.g. [1–4]) and gravitational waves (e.g. [5–12]). Indeed, gravitational microlensing is one of the most important ways of probing compact objects such as "machos" or primordial black holes (PBHs), a dark matter (DM) candidate consisting of compact objects formed in the early Universe. Through surveys of a range of sources, microlensing of such pointlike lenses has been used to constrain the fraction of DM such objects can comprise in a wide range of masses (see e.g. [4]).

It has also been proposed that dark matter can instead be comprised of extended objects, such as boson stars (e.g. [13–16]), axion miniclusters [17], and subhalos [18–22]. Like compact objects, such objects can also bend the light of distant stars. Whether this effect can be probed by a microlensing survey depends on the comparison

[*]miguel.romao@durham.ac.uk
[†]djuna.l.croon@durham.ac.uk

between the object radius and the Einstein radius—the characteristic length scale which is a function of the mass of the dark matter lens and the distance to the light source. The effectiveness of (micro-)lensing then depends on the size of the object compared to the Einstein radius: dilute dark objects, which are transparent to light, are ineffective lenses. Using conservative assumptions about the number of events observed, Refs. [23,24] derived modified constraints on extended dark matter objects.

Interestingly, structures with radii close to the Einstein radius may give distinct microlensing signatures. How the mass is spread within these structures affects the lensing effect. This was demonstrated explicitly for microlensing of various dark matter structures in [23–25]. In objects with a flatter density profile, such as boson stars, the microlensing magnification time series can deviate significantly from that expected from a pointlike lens such as a PBH, for example featuring caustics crossings as can be seen in Fig. 1. These distinguishable features of extended dark lenses can in principle be used to make a positive discovery.

In this work we take the first step towards a positive discovery of a microlensing signature by an extended dark object. To this end we develop an analysis pipeline to search for signatures of extended dark matter objects in time series data. Specifically, we train a histogram-based gradient boosted classifier to identify such features, and describe how these observations can be used to search for dark objects in a range of experiments.

This paper is organized as follows. We first review microlensing by extended lenses and estimate the sensitivity of the OGLE survey to the modified light curves of
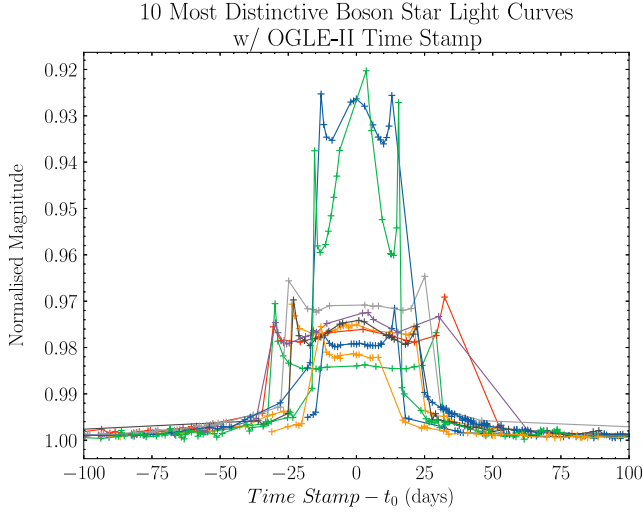
123004-1

FIG. 1. The 10 most distinctive boson star light curves, using the dataset generated with OGLE-II time stamps.

boson stars. We then describe the generation of our dataset and methodology, followed by an analysis of our results and a discussion.

II. MICROLENSING SIGNATURES OF EXTENDED LENSES

Let us first review some basics of gravitational microlensing, largely following the treatment in Ref. [26]. We will need to define some important parameters in the lens setup; a depiction of the geometry can be found in [23]. The observer-lens, lens-source, and observer-source distances are denoted D_L, D_S, and D_{LS} respectively. From the perspective of the observer, the lens center subtends angles of β and θ_i with the source and images of the source respectively. As D_L, D_S, and D_{LS} are much larger than all other scales in the problem, lensing calculations can be simplified using the small angle approximation. In this approximation, the deflections $\alpha = 4GM/(c^2\xi)$ only occur when starlight encounters the "lens plane" perpendicular to the observer-source axis.

Assuming that the lens is spherically symmetric with density distribution $\rho(r)$ [and total mass $M = 4\pi \int_0^\infty dr r^2 \rho(r)$], the lensing equation may be written

$$\beta = \theta - \frac{\theta_E^2}{\theta}\frac{M(\theta)}{M}, \tag{1}$$

with the surface mass density projected onto the lens plane as

$$M(\theta) = 2\pi D_L^2 \int_0^\theta d\theta' \theta' \Sigma(\theta'),$$
$$\Sigma(\theta) = \int_{-\infty}^\infty dz\, \rho\left(\sqrt{D_L^2\theta^2 + z^2}\right), \tag{2}$$

and where [27]

$$\theta_E \equiv \sqrt{\frac{4GM}{c^2}\frac{D_{LS}}{D_L D_S}}, \tag{3}$$

is the pointlike Einstein angle, the value of θ for a pointlike lens $[M(\theta) \to M]$ at zero impact parameter ($\beta = 0$). This in turn defines the pointlike Einstein radius $r_E \equiv D_L \theta_E$ on the lens plane.

As the only way in which the total lens mass M and the distances D_L, D_{LS}, D_S enter the problem is through their contributions to the Einstein radius r_E, it is convenient to express all angles in units of θ_E and all distances r_E. Thus, we define $u \equiv \beta/\theta_E = D_L\beta/r_E$, and $\tau \equiv \theta/\theta_E = D_L\theta/r_E$ (note that the latter was named t in [23,24]), which allows us to rewrite (1) as

$$u = \tau - \frac{m(\tau)}{\tau}, \tag{4}$$

where we have also defined $m(\tau) \equiv M(\theta_E\tau)/M$ which describes the distribution of the lens mass projected onto the lens plane,

$$m(\tau) = \frac{\int_0^\tau d\sigma\sigma \int_0^\infty d\lambda\rho(r_E\sqrt{\sigma^2 + \lambda^2})}{\int_0^\infty d\gamma\gamma^2\rho(r_E\gamma)}, \tag{5}$$

where ρ is the density distribution of the lens.

The lensing equation (4) can be used to find the position(s) of the images θ_i given a position of the lens β. As the images subtend different solid angles than the unlensed source, microlensing alters the observed flux. The magnification is the ratio of the angular extent of an image to that of the source:

$$\mu = \sum_i \left|\frac{\theta_i}{\beta}\frac{d\theta_i}{d\beta}\right| = \sum_i \left|\frac{\tau_i}{u}\frac{d\tau_i}{du}\right|$$
$$= \sum_i \left|1 - \frac{m(\tau_i)}{\tau_i^2}\right|^{-1}\left|1 + \frac{m(\tau_i)}{\tau_i^2} - \frac{1}{\tau_i}\frac{dm(\tau_i)}{d\tau_i}\right|^{-1}. \tag{6}$$

The light curve as a function of time t for a lens with velocity v and minimum impact parameter ξ_{min} can now be found through $\beta r_E = \sqrt{\xi_{min}^2 + v^2 t^2}$. Whether a microlensing event is observable depends on the minimum detectable magnification for a given telescope, as well as the range of cadences at the microlensing survey, which sets the transit timescales to which it is sensitive. Typically, a transit is counted as a lensing event if $\mu > 1.34$, which occurs for a pointlike lens for impact parameter (in units of the Einstein radius) $u \equiv \xi/r_E = \beta/\theta_E = 1$.

In [23,24] the microlensing efficiency of an extended lens compared to that of a pointlike lens was defined as the maximum impact parameter $u_{1.34}$ for which a threshold

magnification is produced: $\mu_{\text{tot}}(u \leq u_{1.34}) \geq 1.34$. For a pointlike lens, $u_{1.34} = 1$ by definition, and the naive expectation for extended lenses would be that $u_{1.34} \leq 1$. Remarkably, this is not necessarily the case: in particular for lenses with a reasonably flat density profile, $u_{1.34}$ can be larger than one. Given $u_{1.34}$, the microlensing differential event rate for a single source with respect to the typical event timescale t_E and $x = D_L/D_S$, observed by a particular experiment, can be calculated as

$$\frac{d^2\Gamma}{dx\,dt_E} = \varepsilon(t_E) \frac{2D_S}{v_0^2 M} f_{\text{DM}} \rho_{\text{DM}}(x) v_E^4(x) e^{-v_E^2(x)/v_0^2}, \quad (7)$$

where $\varepsilon(t_E)$ is the efficiency of telescopic detection, $v_E(x) \equiv 2u_{1.34}(x)r_E(x)/t_E$, $v_0 = 220$ km/s is the dark matter circular speed in the galaxy, and $\rho_{\text{DM}}(x)$ is the DM density projected onto the line of sight, for example following an isothermal profile in the Milky Way galaxy,

$$\rho_{\text{DM}}(r) = \frac{\rho_s}{1 + (r/r_s)^2},$$

$$r \equiv \sqrt{R_{\text{Sol}}^2 - 2xR_{\text{Sol}}D_S \cos \ell \cos b + x^2 D_S^2},$$

with $R_{\text{Sol}} = 8.5$ kpc, $\rho_s = 1.39$ GeV/cm^3, and $r_s = 4.38$ kpc [28]; ℓ and b are the longitude and latitude of the source in galactic coordinates. The total number of events is then given by

$$N_{\text{events}} = N_\star T_{\text{obs}} \int_0^1 dx \int_{t_{E,\text{min}}}^{t_{E,\text{max}}} dt_E \frac{d^2\Gamma}{dx\,dt_E}, \quad (8)$$

where $N_\star$ is the number of observed sources, T_{obs} is the total observation time, and $t_{E,\text{min}}$ ($t_{E,\text{max}}$) is the minimum (maximum) timescale of an event.

Comparing N_{events} with the number of observed events in microlensing surveys, [23,24] set constraints on extended lenses. However, it is important to point out that the surveys identify microlensing events based on a comparison with the pointlike magnification light curve. For extended lenses, in particular lenses with $\tau_m \equiv r_{\text{lens}}/r_E \sim 1$, this is not a good approximation. In Fig. 2 we estimate the range of lens masses and radial sizes for which we expect significant deviations from the pointlike light curve, for the example of a boson star observed in the OGLE-IV survey (5 year dataset) [3], anticipating that the noncompactness of the lens can be resolved for $0.8 < \tau_m < 3$ (which we will verify in the next section). Here we follow the mass profile of boson stars $m(t)$ outlined in the Appendix of [23]. As in this work, the sensitivity is computed using the Poissonian 90% confidence limit based on a signal comprised of dark matter and astrophysical foregrounds (see Ref. [3]). We note that as the OGLE Collaboration did not search for the particular microlensing light curves predicted by boson stars, this is an estimate only. Comparing Fig. 2 with the sensitivity of the
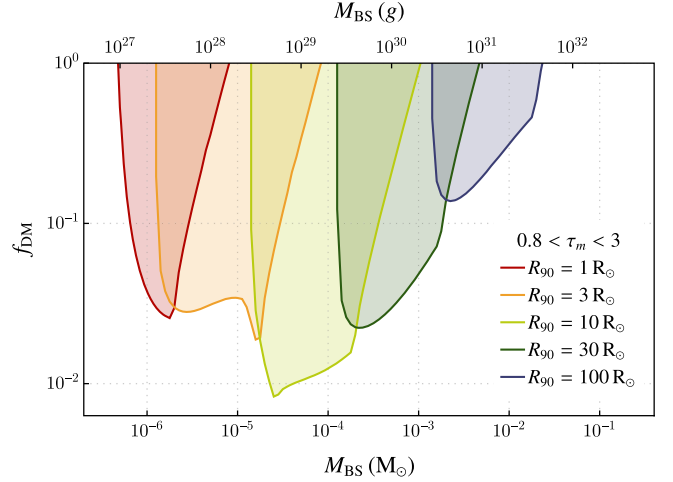

FIG. 2. Sensitivity of the OGLE survey to a modified lensing light curve of a boson star. Here we anticipate that a boson star can be distinguished from a pointlike lens for $0.8 < \tau_m < 3$.

OGLE-IV survey to boson stars in [23], we note that for a given lens size, the lighter masses lead to distinguishable features in the light curve, as can be expected from the dependence of the Einstein radius on M.

Because the Einstein radius varies along the line of sight to the source, for a lens of a particular size r_{lens}, a range of $\tau_m(x)$ is relevant, where $x = D_L/D_S$. One might wonder what the distribution in τ_m is. This is given by $dN/d\tau_m = (dN/dx)(dx/d\tau_m) = f_{\text{DM}}\rho_{\text{DM}}(x)dx/d\tau_m$. For the OGLE sources, we find that this distribution peaks for $\tau_m(x = 0.5)$, and rapidly falls like τ_m^{-3} away from it. Thus, we expect that given a particular lens mass, it is in practise possible to interpret a measurement of τ_m directly in terms of a lens size.

III. DATASET GENERATION AND METHODOLOGY

In this work we adapt MicroLIA, a tool developed and detailed in [29]. The classifier developed in this paper is a machine learning model utilizing the Random Forest algorithm, specifically designed for the detection of microlensing events in astronomical surveys. It is tailored to handle the challenges posed by low-cadence data, which typically suffer from irregular signal sampling and thus lower signal-to-noise ratios, making microlensing event detection more difficult. MicroLIA distinguishes between microlensing and other variable star events using 148 features derived from the light curve and its derivative time series. The classifier categorizes events into classes like microlensing, eclipsing binaries, and regular variable stars, focusing on accurate identification of microlensing amidst these.

In this work, we extend the scope of MicroLIA by including extended dark matter objects which act like nonpointlike microlensing lenses, such as boson stars

(BS, focusing here on non-relativistic condensates of massive bosons without self-interactions) and Navarro-Frenk-White (NFW) subhalos. For this, we generated microlensing light curves for these extended objects using their mass profile, $m(\tau)$, which have previously been calculated in [23]. From these mass profiles, we fit an interpolating function, and for each value of the impact parameter, u, we solve the microlensing equation (4) to obtain the total magnification to the light curve, (6).

The observed impact parameters are closely related to the cadence of a survey. For a lens with a characteristic timescale defined by the crossing of the Einstein radius $t_E \equiv R_E / v$ and a minimum impact parameter u_0 producing a magnification peak occurring at t_0, the values of the observed impact parameters are

$$ u = \sqrt{u_0^2 + \left(\frac{t - t_0}{t_E}\right)^2}, \qquad (9) $$

where t is the survey time. Therefore, the cadence at which the survey collects observational data will have a significant impact in the observed light curve. For this study, we considered two possible cases for the data collection time stamps. In the first case we used OGLE-II time stamps (but not the light curves nor their errors), which are input into MicroLIA and randomly sampled when simulating individual light curves. We note that the regularity of the cadence in OGLE-II is not significantly different from other iterations of the survey. In the second case we considered a *perfectly regular daily cadence*, where time stamps are all equally spaced by the same interval. As we will see below, this case demonstrates some identification opportunities that are obscured by irregular cadences, such as the case with OGLE-II time stamps. We will refer to these to cases as OGLE-II Time stamps and Regular Daily Cadence, respectively.

Both datasets were generated by simulating 100 000 light curves for each of the six classes: cataclysmic variables (CVs), RR Lyrae and Cepheid Variables (VARIABLE), Mira long-period variables (LPV), point-like microlensing (ML), BS, and NFW subhalos. The CV, VARIABLE, LPV, and ML light curves were generated using MicroLIA's simulation, and BS and NFW were generated using our own simulation. We applied the same selection criteria for the three microlensing source events (ML, BS, NFW). The criteria are the same as MicroLIA's, but we further demanded that the observed magnification be at least 1.34, a common criterion imposed by microlensing surveys. The light curves were simulated with a minimum magnitude of 15, a maximum magnitude of 20, and with Gaussian noise.

The extended microlensing sources, BS and NFW, have mass profiles which depend on the parameter τ_m, which follows some distribution depending on the prevalence of such objects. For this study, we sampled $\tau_m \sim \mathcal{U}(0.5, 5)$

logarithmically, since, as discussed in the previous section, this distribution is strongly peaked at the τ_m reached at $x = 0.5$, and the largest deviations in the light curves are expected for $\tau_m \sim 1$. Finally, we sampled the minimal impact parameter, u_0, differently for each of the three microlensing sources

$$ u_0 \sim \begin{cases} \mathcal{U}(0, 1) & \text{for ML} \\ \mathcal{U}(0, 1.5) & \text{for BS} \\ \mathcal{U}(0, 1.1) & \text{for NFW}, \end{cases} \qquad (10) $$

where we emphasize that these are used during the simulation step, i.e. before the selection criteria, including $\mu \geq 1.34$, are imposed.

In total, we generated 600 000 curves, of which only a few $\mathcal{O}(0.1\%)$ with missing values for some of the features were dropped. The dataset was then split into train, validation, and test subsets with proportions 0.5:0.25:0.25, respectively corresponding to around 300 000, 150 000, and 150 000 light curves, with the six classes being equally represented in each of the sets. For each light curve, we computed 74 features using the light curve time series, in addition to the same 74 using the derivative of the light curve time series, to a total of 148 features. The computed features relate to statistics of the time series of the light curve, as well as other time series quantities. See Ref. [29] and MicroLIA's API reference for the full description and reference of each feature.[1] The training set was used to conduct exploratory data analysis and to train machine learning models. The validation set was used for model selection and comparison, and to produce preliminary analysis plots and statistics. The test set was used to produce the final analysis plots and statistics, presented in the next section. The full dataset can be downloaded here [30].

For the multiclassification task, we trained a histogram-based gradient boosted classifier, which is a member of the broader machine learning algorithm family of gradient boosting machines (GBMs). Like random forests (RFs), implemented in MicroLIA, GBMs are ensemble models that leverage the power of multiple weaker estimators, usually small trees, to produce a strong estimator. In RFs the trees are independent of each other and the final prediction is obtained by the ensemble average over all the predictions from the trees in the forest. In GBM the trees are not independent but sequentially trained as to improve on the previous iteration. Schematically, consider $F(X)$ to be the output of the GBM as a function of the data, X. In the first iteration, we want to train a simpler estimator, $F_1(X)$, to match a target label, y. Since the estimator is simple, it will not be very accurate and the prediction will have a certain

---

[1]https://microlia.readthedocs.io/en/latest/autoapi/MicroLIA/features/.

residual error, $y - F_1(X)$. In the second iteration, we train another weak estimator, $h_1(X)$ not on the desired output, but on residual error of the previous step to create a new, improved, estimator $F_2(X) = F_1(X) + h_1(X)$, since in the ideal case $h_1(X) = y - F_1(X)$ and we would have the desired prediction. However, each step, i, will have a residual error $y - F_i(X)$, and so the process can be repeated until the desired accuracy (or maximum number of iterations) is reached. The final GBM output will then be a function of all the steps, $F(X) = F_m(X) = F_{m-1}(X) + h_{m-1}(X)$, where m is the total number of steps. This presentation is schematic and can be generalized to any problem with a differentiable loss function.[2] GBM have been known in the literature to be powerful estimators for tabular data, which we explore in this work. The usual implementation of GBM uses simple trees at each step. However, this requires sorting the data at each iteration, making tree-based GBM computationally heavy for datasets larger than a few tens of thousands examples, which is our case. To mitigate this, a histogram-based variant has been developed that sorts and binarizes (i.e., assign each data point to the bins of the histogram) the data once, producing orders of magnitude speed improvements in training and prediction. In this work, we used the histogram-based GBM implemented by `scikit-learn`'s `HistGradientBoostingClassifier` (HGBC).

In a preliminary study, we compared HGBC against other GBM implementations and scikit-learn's RF, observing all GBM to outperform the RF when computing the area under the curve (AUC) of the receiver operating characteristic (ROC) evaluated on the validation set. We then tuned the hyperparameters of the HGBC, observing minimal improvements of its discriminating performance, for which reason we decided to keep the default hyperparameters for the rest of the study presented herein.[3]

IV. ANALYSIS

In this section, we present the multiclassification analysis on the two generated sets: the first generated with the OGLE-II time stamps, and the second generated with ideal regular daily cadence. The purpose of performing two analyses is to assess whether and how the discrimination power is affected by the cadence.

---

[2]The term *boosting* arises from iteratively boosting the performance using the output of another weak learner. The term *gradient* comes from the observation that fitting successive weak estimators on the residuals of the previous iteration is equivalent to a gradient descent step in a function space spanned by the weak learners.

[3]The performance improvements only impacted at most the third significant digit of the validation ROC AUC, which is arguably at the level of statistical fluctuation of the dataset itself, and therefore meaningless.
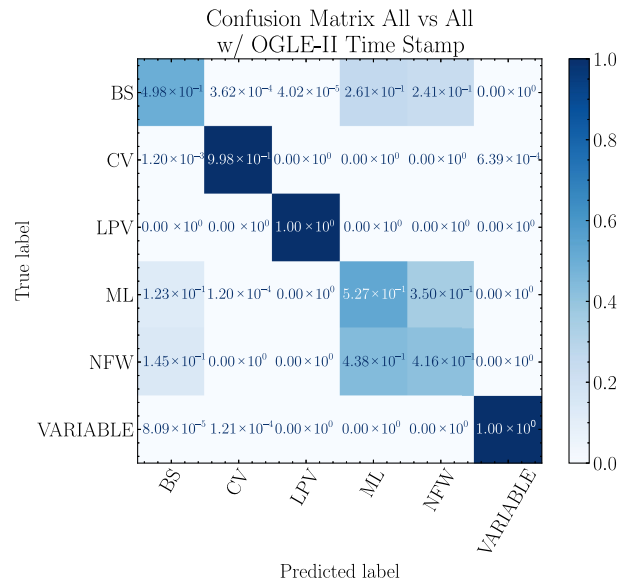

FIG. 3. Confusion matrix for the six-way All vs All multiclassification performed by the HGBC using the dataset generated with OGLE-II time stamps. The entries are rounded to three significant digits.

A. OGLE-II time stamps

In Fig. 3 we present the confusion matrix of the six-way (All vs All) multiclassification performed by the HGBC. We observe that the nonmicrolensing events—CV, LPV, VARIABLE—have minimal to nonexisting overlap with any class. Conversely, the HGBC prediction for the microlensing events—ML, BS, and NFW—has significant overlap. However, we observe that BSs suffer from considerable less contamination from ML and NFW events than these two do of each other. This suggests that BS events are the easiest to isolate, and therefore to detect, of the three microlensing cases.

Since Fig. 3 suggests that BS events are easy to isolate, it is important to understand the nature of the BS we identify. All BSs follow the same mass profile, $m(t)$, which can be wider or narrower, depending on the parameter τ_m, associated with the radius of the BS. In Fig. 4 we show how the probability of a BS being identified as such by the HGBC depends on τ_m. We find that there is a *sweet spot* at $\tau_m \simeq 2$ to maximize the correct identification of BSs, with some BS light curves being classified with high confidence, i.e. with $P(y = \mathrm{BS}|X) \simeq 1$. This also motivates a posterior to our choice of interval $0.8 < \tau_m < 3$ in Fig. 2. In Fig. 1 we present the 10 BS light curves with the highest value of $P(y = \mathrm{BS}|X)$, and we observe that they all exhibit the three peak magnification profile produced by caustics that one expects from BS sources.

Our analysis so far has shown that the discrimination between microlensing sources (ML, BS, NFW) and other sources of light curves (CV, LPV, VARIABLE) is a relatively easy task when OGLE-II cadence is observed,
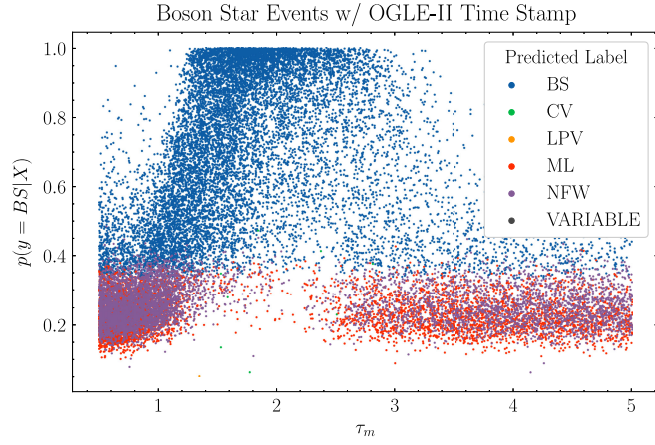
FIG. 4.   Probability of a BS light curve being correctly identified as one by the HGBC vs the boson star τ_m parameter, using the dataset generated with OGLE-II time stamps.
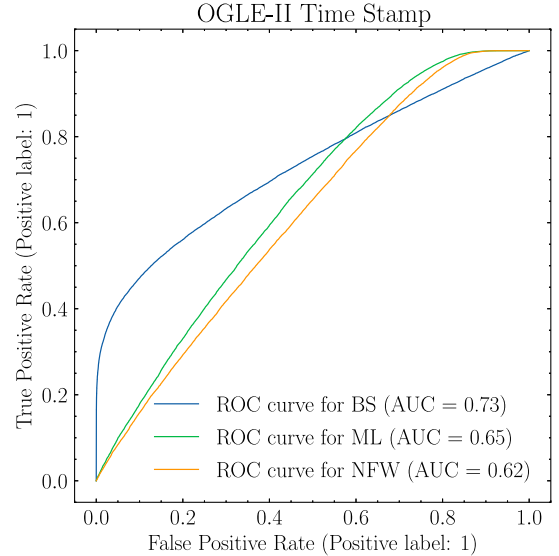

FIG. 5.   ROC curves and their areas obtained from a HGBC trained on the three-way multiclassification task using the dataset generated with OGLE-II time stamps.

with almost nonexisting overlap between these two broad classes. As such, we now focus on the difficult three-way classification focused on discriminating between the three microlensing sources. In Fig. 5 we present the ROC and the AUC of each class (vs the other two) for a HGBC trained on this three-way multiclassification problem.[4] The advantage of analyzing a ROC curve, and its area, over a classification matrix is that the ROC curve captures the classification performance over all possible output threshold cuts, while the confusion matrix only shows the true and false positives of the assigned predicted class as the one that has maximal probability. For example, in Fig. 5 we can see that BS light curves can be isolated with high purity (i.e. with false positive rate around 0.0), while that is not possible for neither ML nor NFW light curves, again reinforcing that BS light curves are easier to identify than the remaining microlensing sources due to their unique three-peak profile. Furthermore, Fig. 5 suggests that pointlike microlensing light curves are only slightly easier to identify than NFW ones, since their ROC curve has a true positive rate greater than that of the NFW for all values of the false positive rates. This is also evidenced by the ML AUC, 0.65, which is greater than that of the NFW, 0.62.

Having been able to isolate the BS light curves from other microlensing sources, we now turn to interpreting our results. Unfortunately, although HBGCs are incredibly powerful estimators they do no provide a clear way of interpreting their prediction, a common challenge when using machine learning in highly dimensional multivariate analyses. In this study, we implemented a backward sequential feature selection (SFS) loop to assess which of the 148 features are relevant for this classification task. The backward SFS loop starts by fitting an HGBC using all
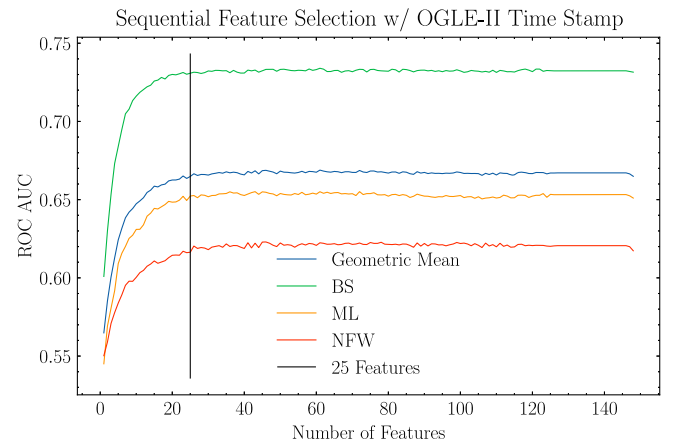

FIG. 6.   Backward SFS (features reduced from right to left) for a HGBC trained on the three-way multiclassification task using the dataset generated with OGLE-II time stamps.

148 features and evaluating its performance on the validation set. Then we train 148 HGBCs on all the possible 148 subsets comprised of 147 features (i.e., with one feature removed) and evaluate their performance on the validation set. Keeping only the best set of 147 features, the same step can be done for the 146 subsets and so one, for a total of 148 iterations totaling 11026 HGBCs, trained on 11026 subset of features, which is a feasible study given the increased training speed offered by the HGBC.[5] In Fig. 6

---

[4]The AUC of a ROC curve can then be seen as an average performance over all possible thresholds set on the predicted probability used to predict the class.

[5]Noting that a set of 148 elements has $2^{148} - 1$ nonempty subsets, 11026 is far fewer than the total possible combinations of features.

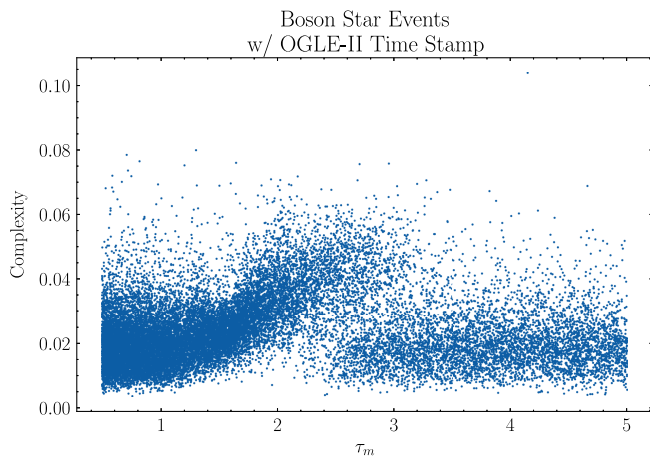FIG. 7. Complexity vs τ_m scatter for BS events with OGLE-II time stamps.
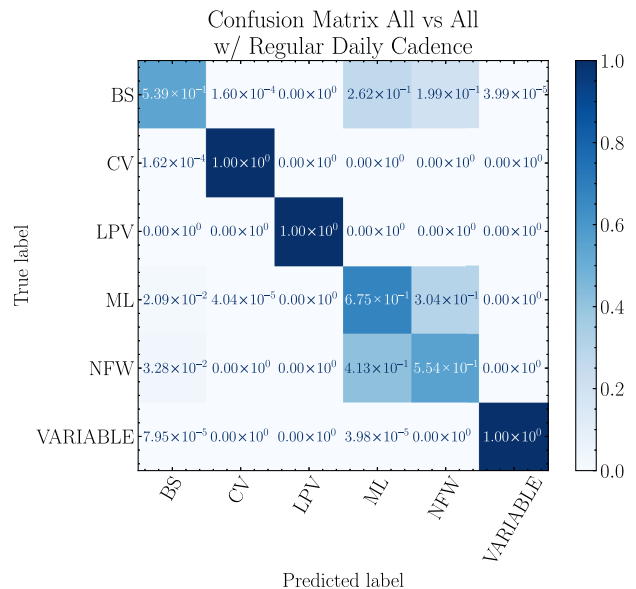

FIG. 8. Confusion matrix for the six-way All vs All multiclassification performed by the HGBC using the dataset generated with regular daily cadence time stamps. The entries are rounded to three significant digits.

we show the performance of the HGBC on the three-way microlensing focused multiclassification for each of the possible sources and their geometric mean. We see that only around 25 features derived from the light curve time series and its derivatives, out of the 148, are relevant for this task.[6] We also note that the discrimination performance degrades for all the three classes at the same stage, suggesting that the final 25 features are relevant for all three classes. The list of the final 25 features by survival ranking is presented in Table I.

For the dataset using OGLE-II time stamps, which is the subject of the analysis in this section, we find that the *complexity* of the time series is the most important feature to separate the light curves from the three microlensing sources, surviving the whole backward selection loop until there is only one. In Fig. 7 we show the distribution of the time series complexity for BS against the τ_m, where we witness a boost in the time series complexity for $\tau_m \simeq 2$, the same sweet spot identified above in Fig. 4, suggesting that the complexity of the time series of a BS light curve is a driving feature for it to be correctly identified as a BS. The same is not observed for the NFW events, which we do not show for the sake of presentation tidiness.

In Fig. 1 we show the 10 most distinctive BS light curves, i.e. the 10 BS light curves with highest $P(y = \mathrm{BS}|X)$ as identified by the HGBC trained on the three-way multiclassification task on the OGLE-II dataset. These light curves all exhibit very clear three magnitude peaks, a hallmark feature arising from the caustics produced by the extended nature of the lens.

---

[6]It is important to point out that while this methodology is aimed at removing uninformative features, it will also remove redundant, i.e. highly correlated, features. Therefore, the final set of features is only unique up to fluctuations of the data and the stochastic initialization of the HGBC. To mitigate this, we fixed the seed of the run, but modifications of the training and validation set could still hold a different set of features.

B. Regular daily cadence time stamps

So far our analysis has focused on light curves simulated using OGLE-II time stamps. This reflects well the sensitivity to extended objects in a realistic microlensing survey, with its given observational constraints. However, one might wonder to what extent the conclusions previously drawn are sensitive to the observation cadence details, especially its irregularity. To address this, we conduct the analysis with light curves simulated with regular daily cadence, i.e. where observations are taken exactly 24 hours apart.

We begin with the six-way All vs All multiclassification task. In Fig. 8 we show the confusion matrix obtained using the HGBC. Although similar to its counterpart with OGLE-II time stamps, Fig. 3, it has some noticeable differences. First, we notice that the contamination in BS positive predictions has decreased from $\mathcal{O}(10\%)$ down to $\mathcal{O}(1\%)$, suggesting an improved capacity of the HGBC in producing a pure sample of BS light curves. Secondly, we notice that the cross misclassification between NFW and ML light curves has also decreased, improving upon the OGLE-II time stamps case.

In Fig. 9 we revisit the probability of a BS light curve being classified as such vs the BS τ_m, but this time for the regular cadence case. We see that the previously identified *sweet spot* at $\tau_m \simeq 2$ has now almost no misclassifications into other classes, showing how a regular cadence can lead to an even better identification of BS light curves; this is due to a better resolution of the light curve being able to capture the three magnification peaks.
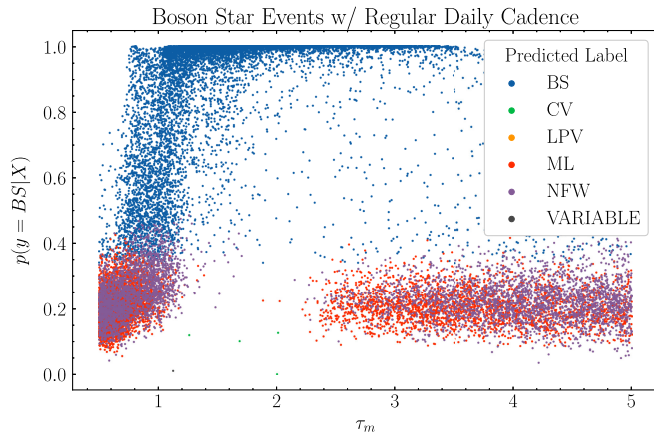
FIG. 9. Probability of a BS light curve being correctly identified as one by the HGBC vs the boson star τ_m parameter, using the dataset generated with regular cadence timestamps.

FIG. 11. Probability of a NFW light curve being correctly identified as one by the HGBC vs the NFW τ_m parameter, using the dataset generated with regular daily cadence time stamps.

As before, we now focus on the three-way multiclassification problem focused on the three microlensing sources. The ROC curves for each of the three classes and their respective AUC can be seen in Fig. 10. Again, we see that there are some noticeable differences to the OGLE-II case shown in Fig. 5. The first thing to notice is that all AUC have increased compared to the OGLE-II time stamps case, indicating an easier discrimination when using regular daily cadence time stamps. Next, we see how the ROC curve for the BS is significantly more peaked for small false positive rate, providing further evidence that the BS light curves are better classified with regular daily cadence. More interestingly, however, is how the ML and NFW ROC curves no
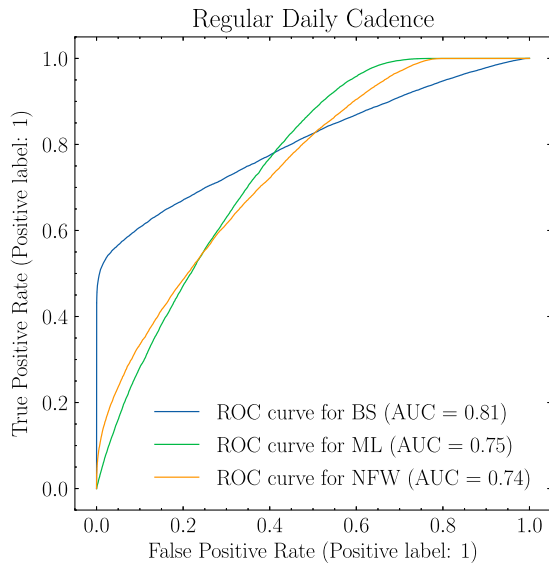

FIG. 10. ROC curves and their areas obtained from a HGBC trained on the three-way multiclassification task using the dataset generated with regular daily cadence time stamps.
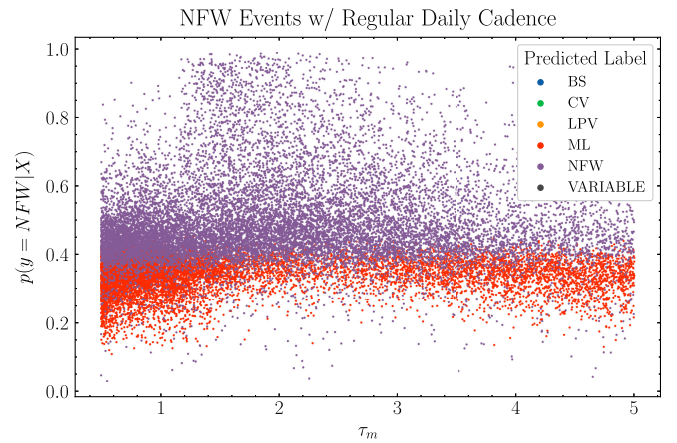
longer follow the same trend as in the OGLE-II case. More precisely, we can observe that they cross, as before they did not. This happens halfway through the curve, with the NFW ROC curve having the *upper hand* over the ML ROC curve for lower values of the false positive rate, with an added feature that the ROC curve for the NFW can have nonvanishing true positive rate at low false positive rate values. This implies that with regular daily cadence it is possible to isolate NFW light curves with little to no contamination of any other class, something that was impossible to achieve using the OGLE-II cadence.

The previous result points at the possibility in principle of completely isolating NFW light curves. It is then important to understand the nature of the NFW light curves that we can isolate. In Fig. 11 we show how the probability for an NFW light curve to be correctly identified as such can vary with the NFW τ_m parameter. Although less noticeable than what we saw before for the BS light curves, we can observe a *sweet spot* for correct classification at $1 \lesssim \tau_m \lesssim 2$, with some light curves being assigned $P(y = \mathrm{NFW}|X) \simeq 1$. We notice that the equivalent scatter for the OGLE-II time stamps, not shown here to declutter the presentation, does not exhibit this pattern, suggesting that there is important information in the light curve that can only be obtained with regular cadence.

Contrary to BS light curves, NFW light curves do not have a clear visual profile compared to pointlike microlensing sources. In Fig. 12 we compare the 100 most easily identifiable pointlike microlensing curves, i.e. those with the highest $P(y = \mathrm{ML}|X)$, against the 100 most easily identifiable NFW curves, i.e. those with the highest $P(y = \mathrm{NFW}|X)$, where the probabilities are obtained from the HGBC trained on the three-way multiclass classification on the regular cadence time stamps dataset. We normalized the magnitudes by min-maxing their values to fall under the [0, 1] interval to aid visual comparison.
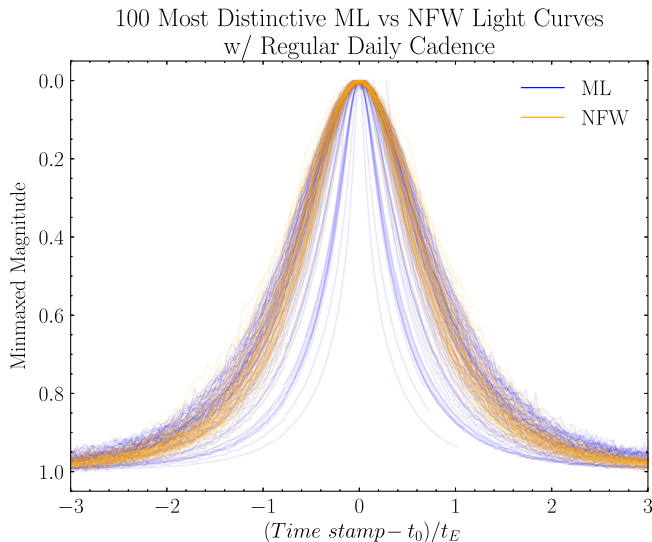
FIG. 12.   Comparison between the 100 most easily identifiable ML curves and the 100 most easily identifiable NFW curves.

Although the difference is very nuanced, the NFW curves tend to be *narrower* than the ML curves. This subtlety explains why regular cadence is so important for identifying NFW light curves, as one needs a better *resolution* of the light curve profile to be able to identify this nuance, which proved impossible when using the OGLE-II time stamps.

Finally, we perform a backward SFS for the regular daily cadence dataset using a three-way HGBC focused on the microlensing classes. The discrimination performance vs the number of features is shown in Fig. 13, where we observe a similar trend than that already discussed for the OGLE-II time stamps dataset with the ROC AUC degrading significantly below 25 features. However, it is worth noting

that the performance degrades first for the NFW and ML cases, with the ROC AUC associated with the classification of the BS light curves staying at its maximum value until around 15. This suggests that there are less relevant features to distinguish BS light curves from the other classes than there are to correctly identify the others. In Table I we present the top 25 features, and we observe considerable overlap with the OGLE-II study. Perhaps curiously, we count 10 of the features to have been computed on the derivative of the time series, whereas for the OGLE-II the number is six, possibly hinting at the importance of the derivative of the time series in identifying the nuanced differences in shape between the ML and the NFW light curve magnification peak.

V. DISCUSSION

In this work we have studied the observation of microlensing signatures due to extended dark objects in time series data, and its distinction from other signals, most importantly pointlike lenses. We have focused on two lens profiles, each exemplifying a class of objects: the profile of a NFW subhalo is more peaked, whereas boson stars have a more diffuse profile. As expected, the boson stars can be confidently distinguished from pointlike lenses, for $0.8 \lesssim \tau_m (= r_{\text{lens}}/r_E) \lesssim 3$. This is owed to the characteristic caustics in the light curves for these objects.

As an exercise, we studied how the regularity of the time series cadence affects the confidence of the detections. Interestingly, for regular daily cadence we also find confident detections of NFW subhalos. These detections occur for $0.9 \lesssim \tau_m \lesssim 4$, despite the fact that no caustics are observed in the light curves. Though the regularity of (daily) observations is dependent on many conditions that are beyond the observer's control, this is an interesting observation which warrants investigation for other microlensing scenarios and targets.

Stellar binaries or exoplanets orbiting a lensing body may give rise to perturbations and caustics in the light curve. Unlike for extended dark objects, these are asymmetric or one-sided features (see e.g. [31]). Microlensing has allowed for the discovery of over 100 exoplanets, particularly of near-terrestrial size [32]. In this work, we have not considered these signatures, which may give rise to confusion with boson star identifications with low significance. We will leave such an analysis for future work.

We performed the analysis in this work by extending the MicroLIA algorithm, which utilizes 148 features derived from the light curves in a single band to distinguish between objects. Our SFS analysis showed that of these, only 25 were needed for the optimal classification. For our regular cadence analysis, further features could help distinguish between NFW subhalos and pointlike lenses. In future work, we will study the potential benefits of learning directly on the light curves.
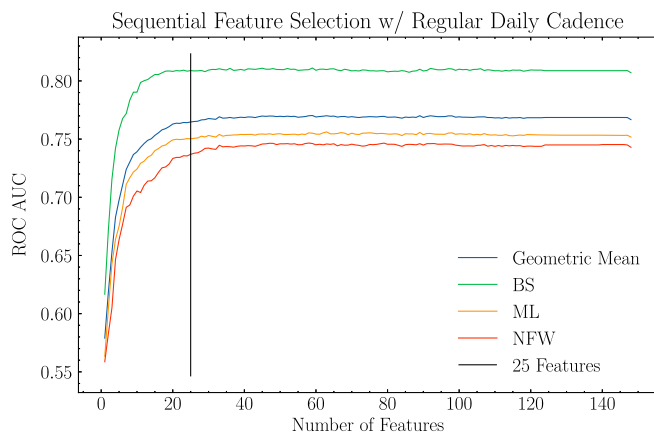

FIG. 13.   Backward SFS (features reduced from right to left) for a HGBC trained on the three-way multiclassification task using the dataset generated with regular daily cadence time stamps.

We note that in our analysis we have assumed a Gaussian noise model, which may not be realistic. In future work towards the application of our methodology, further noise models should be considered. These could lead to larger misclassification between classes of events.

Microlensing surveys typically only release data on candidate events, which have passed through a strict selection in which our extended dark object light curves were likely cut. Exceptions include the UKIRT Microlensing Survey [33,34] and the VISTA Variables in the Via Lactea Survey [35], which we plan to analyze in future work. Upcoming microlensing opportunities include the launch of infrared astronomy experiments in the mid 2020s. In particular, the Nancy Grace Roman Space Telescope (previously WFIRST; a space mission by NASA) has as a key objective to discover exoplanets through the microlensing technique [36], but will also probe primordial black holes [37] and extended dark objects [38].

Finally, for some surveys, the finite extent of the source becomes important. This is the case, for example, for the Subaru-hsc survey of M31 [39]: because of its sensitivity to small transit times (and hence small Einstein radii), the angular extent of source stars corresponds to a distance at the lens larger than the Einstein radius, suppressing the magnification relative to pointlike sources [16,40–42]. We leave an analysis of the magnification curves with finite source effects for future work.

VI. SOFTWARE

For the computation of the mass profile of the extended sources and the sensitivity estimation, we used *Mathematica* version 12.

The dataset generation, machine learning training, and the final analysis were performed in PYTHON 3.9.18, making use of several packages, of which we make especial note:

(i) Light curve simulation and time series feature extraction was performed using MicroLIA. We used the code directly from the MicroLIA github repository, as it includes considerable changes to the code provided with the packaged version 2.6.0. More precisely, the version used makes proper use of the derivative time series, as well as its errors as statistical weights. These two fixes were contributions of this work.

(ii) `scikit-learn` 1.3.2 for HGBC implementation.

(iii) `mlxtend` 0.23.0 for the backwards SFS.

(iv) `SciencePlots` 2.1.1 for plotting.

The full list of packages and their versions can be found in the `requirements.txt` file included in the code repository hosting the code used in this work.[7]

APPENDIX: BACKWARD SEQUENTIAL FEATURE SELECTION RESULTS

In Table I we collect the top 25 features obtained by the backward SFS, for both the OGLE-II and the regular daily cadence time stamps datasets. The details of the features can be seen in [29] and in MicrioLIA's code. The feature rank is associated with how long it survived the SFS loop. For example, a rank 1 means that the feature is present in all subsets down to the last subset of a single feature, whereas a rank n means that the feature is present up until the subset of size n, but is then removed and is not present in smaller subsets. This set of features is only unique up to fluctuations of the data and the stochastic variability associated to HGBC training. In principle, any feature presented in this table could be replaced by a highly correlated counterpart, and as such the list presented herein should be interpreted as the maximal set of mutually uninformative features obtained using our training and validation sets.

---

[7]https://gitlab.com/miguel.romao/microlensing-extended-objects-machine-learning.

TABLE I. Top 25 features, ranked by *survivability* of the backward sequential feature selection loop. Features computed in the derivative of the time series are identified as such. See Ref. [29] for details about the features.

Feature rank	OGLE-II	Daily regular cadence
1	Complexity	StetsonJ (derivative)
2	Median buffer range	Half mag amplitude ratio
3	Ratio recurring points	FluxPercentileRatioMid80
4	Amplitude	Median buffer range
5	MedianAbsDev	Count below
6	FluxPercentileRatioMid80	MedianAbsDev
7	Count below	Check max last loc
8	Cusum	Longest strike below
9	Shapiro wilk	Complexity
10	Mean change	PercentDifferenceFluxPercentile
11	Longest strike above (derivative)	Longest strike above
12	Mean second derivative	Shapiro wilk (derivative)
13	Number of crossings	FluxPercentileRatioMid65
14	LinearTrend	Number of crossings
15	Longest strike below (derivative)	Number cwt peaks
16	Sample entropy	Number cwt peaks (derivative)
17	FluxPercentileRatioMid65	Time reversal asymmetry (derivative)
18	Mean change(derivative)	FluxPercentileRatioMid35
19	Longest strike above	Peak detection(derivative)
20	Mean n abs max (derivative)	Mean n abs max (derivative)
21	FluxPercentileRatioMid20	Longest strike above (derivative)
22	Quantile (derivative)	Check min last loc
23	Check min last loc	Quantile (derivative)
24	Gskew	Sample entropy (derivative)
25	Check max last loc (derivative)	Below3 (derivative)

---

[1] C. Alcock *et al.*, Astrophys. J. **499**, L9 (1998).

[2] H. Niikura *et al.*, Nat. Astron. **3**, 524 (2019).

[3] H. Niikura, M. Takada, S. Yokoyama, T. Sumi, and S. Masaki, Phys. Rev. D **99**, 083503 (2019).

[4] A. M. Green and B. J. Kavanagh, J. Phys. G **48**, 043001 (2021).

[5] S. Bird, I. Cholis, J. B. Muñoz, Y. Ali-Haïmoud, M. Kamionkowski, E. D. Kovetz, A. Raccanelli, and A. G. Riess, Phys. Rev. Lett. **116**, 201301 (2016).

[6] Y. Ali-Haïmoud, E. D. Kovetz, and M. Kamionkowski, Phys. Rev. D **96**, 123523 (2017).

[7] B. J. Kavanagh, D. Gaggero, and G. Bertone, Phys. Rev. D **98**, 023536 (2018).

[8] G. Bertone *et al.*, SciPost Phys. Core **3**, 007 (2020).

[9] B. P. Abbott *et al.* (LIGO Scientific and Virgo Collaborations), Phys. Rev. Lett. **123**, 161102 (2019).

[10] Z.-C. Chen and Q.-G. Huang, J. Cosmol. Astropart. Phys. 08 (2020) 039.

[11] G. Franciolini, Primordial black holes: From theory to gravitational wave observations, Ph.D. thesis, Université de Genève, 2021.

[12] D. Croon, S. Ipek, and D. McKeen, Phys. Rev. D **107**, 063012 (2023).

[13] R. Ruffini and S. Bonazzola, Phys. Rev. **187**, 1767 (1969).

[14] M. Gleiser, Phys. Rev. D **38**, 2376 (1988); **39**, 1257(E) (1989).

[15] F. E. Schunck and E. W. Mielke, Classical Quantum Gravity **20**, R301–R356 (2003).

[16] D. Croon, J. Fan, and C. Sun, J. Cosmol. Astropart. Phys. 04 (2019) 008.

[17] M. Fairbairn, D. J. E. Marsh, and J. Quevillon, Phys. Rev. Lett. **119**, 021101 (2017).

[18] A. L. Erickcek and K. Sigurdson, Phys. Rev. D **84**, 083503 (2011).

[19] G. Barenboim and J. Rasero, J. High Energy Phys. 04 (2014) 138.

[20] J. Fan, O. Özsoy, and S. Watson, Phys. Rev. D **90**, 043536 (2014).

[21] J. A. Dror, E. Kuflik, B. Melcher, and S. Watson, Phys. Rev. D **97**, 063524 (2018).

[22] P. W. Graham, J. Mardon, and S. Rajendran, Phys. Rev. D **93**, 103520 (2016).

[23] D. Croon, D. McKeen, and N. Raj, Phys. Rev. D **101,** 083013 (2020).

[24] D. Croon, D. McKeen, N. Raj, and Z. Wang, Phys. Rev. D **102,** 083021 (2020).

[25] Y. Bai, A. J. Long, and S. Lu, J. Cosmol. Astropart. Phys. 09 (2020) 044.

[26] R. Narayan and M. Bartelmann, in 13th Jerusalem Winter School in Theoretical Physics: Formation of Structure in the Universe Jerusalem, Israel, 1996 (1996), arXiv:astro-ph/9606001.

[27] A. Einstein, Science **84,** 506 (1936).

[28] M. Cirelli, G. Corcella, A. Hektor, G. Hütsi, M. Kadastik, P. Panci, M. Raidal, F. Sala, and A. Strumia, J. Cosmol. Astropart. Phys. 03 (2011) 051.

[29] D. Godines, E. Bachelet, G. Narayan, and R. Street, Astron. Comput. **28,** 100298 (2019).

[30] M. Crispim Romão and D. Croon, Light curves for variable, point-like microlensing, and extended objects microlensing sources with regular cadence and OGLE-II timestamps cadence (2024), 10.5281/zenodo.10566869.

[31] S. Khakpash, J. Pepper, M. Penny, B. S. Gaudi, and R. Street, Astron. J. **161,** 132 (2021).

[32] Nasa exoplanet archive, https://exoplanetarchive.ipac.caltech.edu/ (accessed: 25-01-2024).

[33] Ukirt microlensing survey information, https://exoplanetarchive.ipac.caltech.edu/docs/UKIRTMission.html (accessed: 25-01-2024).

[34] Y. Shvartzvald, G. Bryden, A. Gould, C. Henderson, S. Howell, and C. Beichman, Astron. J. **153,** 61 (2017).

[35] D. Minniti, P. Lucas, J. Emerson, R. Saito, M. Hempel, P. Pietrukowicz, A. Ahumada, M. Alonso, J. Alonso-Garcia, J. I. Arias *et al.*, New Astron. **15,** 433 (2010).

[36] D. Spergel *et al.*, arXiv:1503.03757.

[37] J. Fardeen, P. McGill, S. E. Perkins, W. A. Dawson, N. S. Abrams, J. R. Lu, M.-F. Ho, and S. Bird, arXiv:2312.13249.

[38] W. DeRocco, N. Smyth, and V. Takhistov, Astrophys. J. Lett. **965,** L3 (2024).

[39] H. Niikura *et al.*, Nat. Astron. **3,** 524 (2019).

[40] H. J. Witt and S. Mao, Astrophys. J. **430,** 505 (1994).

[41] P. Montero-Camacho, X. Fang, G. Vasquez, M. Silva, and C. M. Hirata, J. Cosmol. Astropart. Phys. 08 (2019) 031.

[42] N. Smyth, S. Profumo, S. English, T. Jeltema, K. McKinnon, and P. Guhathakurta, Phys. Rev. D **101,** 063005 (2020).