Received: 30 May 2021 | Revised: 18 October 2021 | Accepted: 20 October 2021

DOI: 10.1002/spe.3052

SPECIAL ISSUE PAPER

WILEY

Dynamics signature based anomaly detection

Ivan Hendy Goenawan[1] | Zhihui Du[2] | Chao Wu[3] | Yankui Sun[1] | Jianyan Wei[3] | David A. Bader[2]

[1]Department of Computer Science and Technology, Tsinghua University, Beijing, China

[2]Department of Data Science, New Jersey Institute of Technology, Newark, New Jersey, USA

[3]National Astronomical Observatories, Chinese Academy of Sciences, Beijing, China

Correspondence
Zhihui Du, Department of Data Science, New Jersey Institute of Technology, Newark, NJ 07102, USA.
Email: zhihuidu@gmail.com
Chao Wu, National Astronomical Observatories, Chinese Academy of Sciences, Beijing 100012, China.
Email: cwu@nao.cas.cn

Funding information
Key Research and Development Program of China, Grant/Award Number: 2016YFB1000602; National Science foundation, Grant/Award Number: 2109988; MOE Research Center for Online Education Foundation, Grant/Award Number: 2016ZD302; National Natural Science Foundation of China, Grant/Award Numbers: 11690023, 61762074, 61073008, 61272087, 61440057

Abstract

Identifying anomalies, especially weak anomalies in constantly changing targets, is more difficult than in stable targets. In this article, we borrow the dynamics metrics and propose the concept of dynamics signature (DS) in multi-dimensional feature space to efficiently distinguish the abnormal event from the normal behaviors of a variable star. The corresponding dynamics criterion is proposed to check whether a star's current state is an anomaly. Based on the proposed concept of DS, we develop a highly optimized DS algorithm that can automatically detect anomalies from millions of stars' high cadence sky survey data in real-time. Microlensing, which is a typical anomaly in astronomical observation, is used to evaluate the proposed DS algorithm. Two datasets, parameterized sinusoidal dataset containing 262,440 light curves and real variable stars based dataset containing 462,996 light curves are used to evaluate the practical performance of the proposed DS algorithm. Experimental results show that our DS algorithm is highly accurate, sensitive to detecting weak microlensing events at very early stages, and fast enough to process 176,000 stars in less than 1 s on a commodity computer.

KEYWORDS

anomaly detection, dynamics features, gravitational microlensing, periodic variable stars, time series

1 | INTRODUCTION

Anomaly detection[1] methods have been widely used in astronomical data analysis. However, identifying an anomaly event completely hidden in the valley of a periodic variable star's light curve is a challenging problem. The difficulty lies in that distinguishing the abnormal change from the normal behavior of a variable star is not easy. When we measure the position of a given periodic variable star in the brightness space, we may borrow the dynamics metrics (displacement, velocity, and acceleration) to describe a periodic variable star's normal behaviors. Such dynamics metrics are named as the dynamics signature (DS) of a variable star in multi-dimensional feature space in this article. Because the dynamics states of an abnormal motion and a normal motion in the multi-dimensional feature space are very different, it will be easy to identify the abnormal motion from normal ones. We build DS to capture the track of any continuous normal

behaviors or normal motions. Once the DS is constructed, any dynamics state that cannot fit in the track will be regarded as an anomaly.

Under our scenario, a two-dimensional feature space that includes displacement and velocity dimensions is enough to capture a variable star's dynamics motion in one-dimension brightness space. A dynamics criterion, which means whether the distance between a star's current dynamics state is beyond a given threshold compared with the given DS, is proposed to identify the abnormal changes. The advantage of the proposed DS method is that it can discover abnormal changes that happened within the normal brightness range, which are hard to identify with existing brightness threshold-based methods. DS method is very sensitive to weak anomalies in brightness (less than 3-sigma threshold). We can employ the DS method to significantly reduce the false positive rate with high detection sensitivity.

Microlensing is a typical anomaly and the detection of gravitational microlensing events can aid in the discovery of galactic structure,[2] planet,[3–5] exoplanet,[6] blackhole,[7] with a broad range of masses and with distances up to thousands of light-years away from the earth.[8,9] In this article, we will employ microlensing to design and evaluate our anomaly detection method.

One of the leading teams in microlensing detection is OGLE[10] which is based at the Las Campanas Observatory, Chile. MACHO[11] and EROS[12] are two other important team efforts in microlensing. There are many new microlensing projects, such as KMTNet,[13] MOA.[14] All these projects advance microlensing research significantly.

A highly efficient anomaly detection algorithm that can be used in large scale sky survey data generated by wide-field high cadence telescopes is critical to significantly prompt the research in time-domain astronomy. This work focuses on real-time analysis of short-timescale anomaly in light curves based on the super-large field of view optical instruments. The ground-based wide-angle camera system (GWAC) is an array of 40 cameras designed to observe in the visible domain the prompt optical emission of gamma-ray bursts (GRBs). It is part of the ground segment for the SVOM (Space-based multi-band astronomical Variable Objects Monitor) mission.[4,5,15,16] Although its primary mission is to observe GRB prompt emissions, the high-cadence nature of GWAC also makes it an ideal instrument for detecting transient events such as short-timescale gravitational microlensing events.

Given a large number of stars (about 176,000) that each GWAC camera will be observing and the short time (15 s) with which it produces new images, a sufficiently fast algorithm for detecting microlensing events at their early stages in real-time is another great challenge. Besides, the false positive rate should ideally be very close to zero to improve the efficiency of the follow-up telescopes. We target all these challenging problems in this article.

The major contributions of this article are as follows.

- The concept of DS is proposed to identify the anomaly in a periodic variable star. The DS built in multi-dimensional feature space can clearly distinguish the abnormal changes even within the normal brightness range.
- A highly optimized ***DS*** (dynamics signature) algorithm is presented to show how to implement this idea to detect short-timescale microlensing events in millions of periodic variable stars in real-time.
- Experimental results show that the proposed method is highly accurate to detect weak microlensing events at very early stages, and fast enough to process millions of stars in an online way under a practical scenario.

The rest of the article is organized as follows. In Section 2 we propose the basic idea of our method and formulate the question as a DS based anomaly identification problem. The corresponding DS algorithm description and detailed discussion are presented in Section 3. Two datasets are used to evaluate our method from different aspects and they are described in Section 4. We evaluate our method with a large number of light curves from different aspects and the experimental results are given in Section 5. The related research is presented in Section 6 and we conclude our work in Section 7.

2 | PROBLEM FORMULATION

The essential idea proposed in this article is the concept of DS which can be built in multi-dimensional dynamics feature space to capture the anomaly change of a variable star in brightness space.

To identify an anomaly such as a microlensing event, we need to distinguish the abnormal microlensing effect from the normal brightness behaviors of a variable star. The superposition of microlensing and normal behavior will often fall

into the brightness threshold (3-sigma is often used),[13] so all the brightness threshold-based methods cannot identify the microlensing event hidden in the valley of a variable star's light curve.

We employ the dynamics metrics to analyze different anomalies and find that the motion of a periodic variable star will have a significant change in multi-dimensional feature space if an anomaly event happens. This finding inspires us to propose the concept of the DS of a periodic variable star to identify anomalies, such as microlensing events.

2.1 | Dynamics signature

For a given star, its brightness value at time t can be expressed as $m(t)$. The first-order differential of $m(t)$, $m'(t)$, which means the change of brightness at time t, is another essential feature that shows the dynamics behavior of a variable star. The higher-order differential of $m(t)$ can also be used if necessary. Here in this article, we employ $m(t)$, $m'(t)$ that are often called *displacement* and *velocity* in dynamics metrics as the basic features to build the DS of a periodic variable star.

For the same brightness value, the corresponding value in brightness change may be very different, and vice versa. Since both $m(t)$ and $m'(t)$ can be used to investigate the dynamics behaviors of a variable star from different views, in this article we combine them together as a tuple $<m(t), m'(t)>$ to be one point in a two-dimensional feature space. All such points will form the complete DS of a given star. A DS is the track of a variable star in a two-dimensional feature space.

If the period of a variable star is T, we will have $m(t) = m(t+T)$ and $m'(t) = m'(t+T)$. So the complete ***DS*** can be expressed as a set with the formula below.

$$\mathrm{DS} = \left\{ < m(t), m'(t) > | t_0 \le t < t_0 + T \right\} \tag{1}$$

where t_0 is any observation time point. Under practical scenarios, ***DS*** can be built from the normal history data which covers at least one period T.

2.2 | Dynamics distance and threshold

We propose the concept of dynamics distance to measure the Euclidean distance between a given dynamics state or point and a DS in the dynamics feature space. The larger the distance is, the higher possibility the given dynamics state is an outlier. So for a given dynamics state $Y =< m(t_Y), m'(t_Y) >$, we will check if Y is far away enough from DS to decide it is an outlier or not. For simplicity, in Figure 1 the elliptic curve stands for a given ***DS*** and the red points stand for some anomalies that are far away from the ***DS***.

For any dynamics state $Y =< m(t_Y), m'(t_Y)$, let $X =< m(t_X), m'(t_X) > \in$ DS be any dynamics state of a given ***DS***. The distance Dist(Y, DS) between Y and DS is defined as follows.

$$\mathrm{Dist(Y, DS)} = Min \left\{ \sqrt{(m(t_Y) - m(t_X))^2 + (m'(t_Y) - m'(t_X))^2} | \forall \mathrm{X} \in \mathrm{DS} \right\} \tag{2}$$

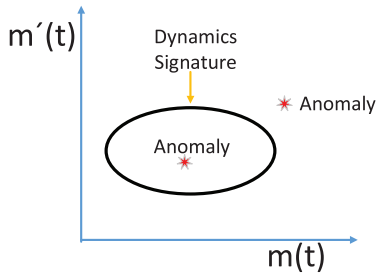


FIGURE 1 Schematic diagram to show how to identify anomaly events using the DS. The farther a point is from the DS, the more likely it is an anomaly

Yet, this distance is calculated based on the absolute values of the brightness and its changing speed. We cannot use the same dynamics threshold value to measure the relative dynamics changes of different periodic variable stars. To solve this problem, we assume that both *m(t)* and *m'(t)* will follow the normal distribution. Thus we can calculate the mean (μ) and SD (σ) of a time series and then transform any variable X into $Z = (X\text{-}\mu)/\sigma$ following the standard normal distribution. In this way, we can use the same normalized dynamics threshold for different periodic variable stars. We use z_*m(t)* and z_*m'(t)* to stand for the normalized results of *m(t)* and *m'(t)*. So the normalized distance from any dynamics state Y to DS can be defined as follows.

$$\text{NormalizedDist(Y, DS)} = Min\left\{\sqrt{(z\_m(t_Y) - z\_m(t_X))^2 + (z\_m'(t_Y) - z\_m'(t_X))^2} | \forall \text{X} \in \text{DS}\right\} \tag{3}$$

Given the dynamics threshold DT, if NormalizedDist(Y, DS) > DT, we will mark Y as an abnormal dynamics state.

Here we express light curve *m(t)* as a continuous function. Under a practical scenario, we will approximate it with discrete points. For microlensing detection, only when we detect several continuous anomaly points, we will give an alert.

3 | DS ALGORITHM DESIGN

In this section, we will present the detailed design on how to develop the DS-based algorithm to identify the weak anomalies. First, we should know how to build the DS using online observation data. Then a quick calculation method to measure the distance between a dynamics state and the corresponding DS is given. At last, we should know how to set up a suitable dynamics threshold to distinguish anomalous events from normal behaviors.

3.1 | Generating the DS

Most variable stars have relatively simple light curves, with only one or at most two ascending and descending phases during one oscillatory period.[17] Each magnitude value will therefore be associated with a few distinct possible rates of change (slopes). In fact, most magnitude values will only be associated with one positive and one negative slope, corresponding to the phase when the star is becoming dimmer and brighter, respectively.

Since we can only get discrete points in practical calculations, we will use Δm to replace or approximate *m'(t)* and employ $<m, \Delta m>$ to build the DS of a given variable star. One DS summarizes the progression of a given variable star changing from one dynamics state to the next.

If we generate a DS showing the relationship between Δm and m using past historical data from a particular star, it will have a roughly elliptical shape. Any deviation from the expected track can be used as a reliable indicator that an anomaly has occurred.

In Figure 2, three examples are given to show how some microlensing events can be hidden into the valley of a variable star and how the anomalous change can be clearly identified in the two-dimensional dynamics feature space. The detailed procedures of the algorithm are given below.

The first step in the DS algorithm is to generate the DS, a 'magnitude vs slope' plot here from historical data that can be used as a reference to determine whether new data points are anomalous. The reference plot can be regenerated every day during the daytime when the telescope is not operating. In the current implementation, we use data points obtained in days from CurrentDay-32 to CurrentDay-2. Assuming around 8 h worth of observation per day and a refresh rate of 15 s, around 58,000 (30 days * 1920 observations) data points will be used for the reference plot for each star. We found that this provides a good balance between coverage and memory usage. Note that data points from the most recent 2 days are excluded to prevent anomalous observations to be included in the reference plot.

To obtain the approximate instantaneous slope at each data point, linear regression is performed on all data points generated in the 15 min period prior to it. Also, the average magnitude within the time window is also calculated. This pair of values (m, Δm) becomes a single point in the reference 'm vs Δm' plot. The same process is repeated for all remaining historical data points.

The window size of 15 min is chosen according to the timescale of the anomaly that we are interested to detect. Based on our experience, a window size of 15 min works well for identifying anomalies with timescales ranging from 1 h to around 1 day, which is the main focus of the current study. If the window size chosen is too short, the method will be too

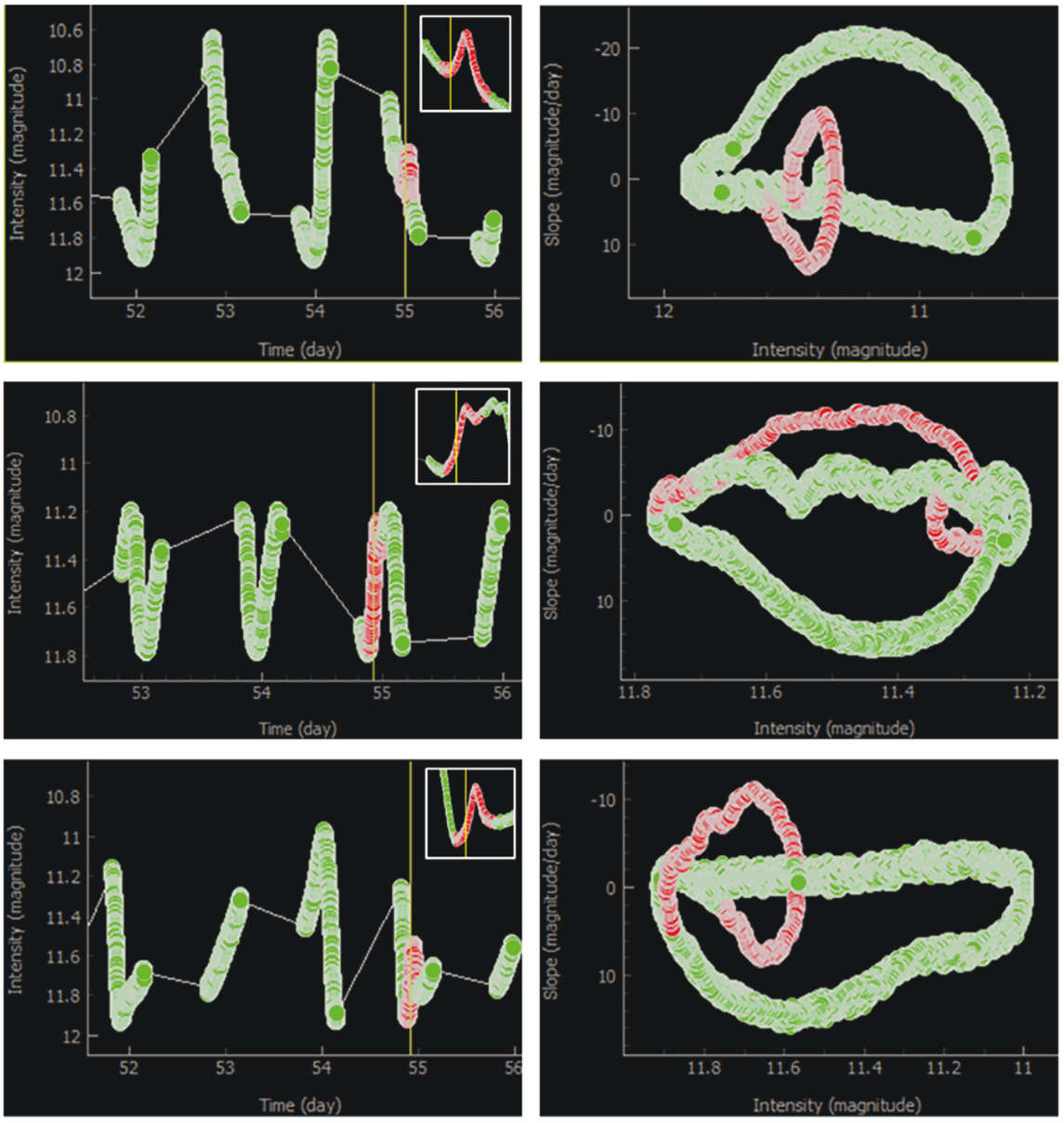


FIGURE 2 Three examples of simulated light curves of variable stars (left) with weak microlensing signals present (shown in red), along with their corresponding 'm vs Δm' plots (right). The gaps in the light curves signify the intervals when the stars are not observed. The yellow vertical lines show the anomaly alert time of our DS algorithm. The 'm vs Δm' reference curves are the DSs of different variable stars. Anomalous signals will manifest themselves as deviations from the reference curves (red in right)

sensitive to random fluctuations caused by noise. On the other hand, anomalies that have shorter timescales would likely go undetected if the window size chosen is too long.

3.2 | Calculating the dynamics distance

We use the GWAC system as an example of how to calculate the dynamics distance. GWAC takes a new image approximately every 15 s. When the latest magnitude reading for a particular star arrives, the corresponding m and Δm values will be calculated from the preceding 15-min window using the same method described in the previous section. Since our DS algorithm uses the latest 15 min of data to produce each (m, Δm) pair, real-time anomaly detection only begins 15 min after the telescope is first turned on. The number of data points in the 15-min window does not have to be constant, so

the algorithm can tolerate missing observations. However, we can set a minimum number of data points that need to be present in each 15-min window (currently 50) for the resulting (m, Δm) values to count as a valid observation.

The DS algorithm then measures how much the new observation deviates from historical data by calculating the average Euclidean distance between the new (m, Δm) point and its 20 nearest neighbors in the reference plot or DS. In order to do this efficiently, we developed a simple heuristic approach for identifying these closest neighbors. This involves transforming the in-memory representation of the reference curve using the steps described below. The procedure can be performed right after the reference plot is generated during maintenance time.

All m and Δm values are first normalized by dividing them with $m_{\max} - m_{\min}$ and $(\Delta m)_{\max} - (\Delta m)_{\min}$, respectively (here we employ a simple normalization method). This has the effect of equalizing the scales in the x and y-axis, so that deviations in both m and Δm have equal contributions to the calculated deviation metric. Since the reference plots generated from stars' light curves have roughly elliptical shapes, the reference plot is then partitioned in a circular manner into 360 slices of equal angular widths along its midpoint, defined simply as the point with the coordinate $(\frac{x_{\min}+x_{\max}}{2}, \frac{y_{\min}+y_{\max}}{2})$. The coordinate of each point is then converted into a polar coordinate which specifies a) which slice it belongs to and b) its distance from the midpoint. Each slice is represented with separate arrays containing the distances from the midpoint to the individual data points that fall in its area. Each slice is then individually sorted (see Figure 3). Since each historical data point is now represented with only a single number, this approach also significantly reduces memory usage.

When a new data point arrives, its (m, Δm) coordinate can be quickly converted into a polar coordinate. The angular component is then used to find the corresponding slice. Meanwhile, the radial component is used to find the 20 nearest neighbors in the sorted array. This can be done efficiently using binary search.

3.2.1 | Corner cases

In some situations, the generated reference curve may not have a nice circular shape. First, it is possible that the reference dataset does not cover all phases of the star's oscillation. This can occur, for example, when the period of the star happens to be a multiple of 1 day so that some phases can never be observed during night time. It is therefore possible that some slices in the boundary regions contain less than 20 historical data points (Figure 4A). When a new observation falls into

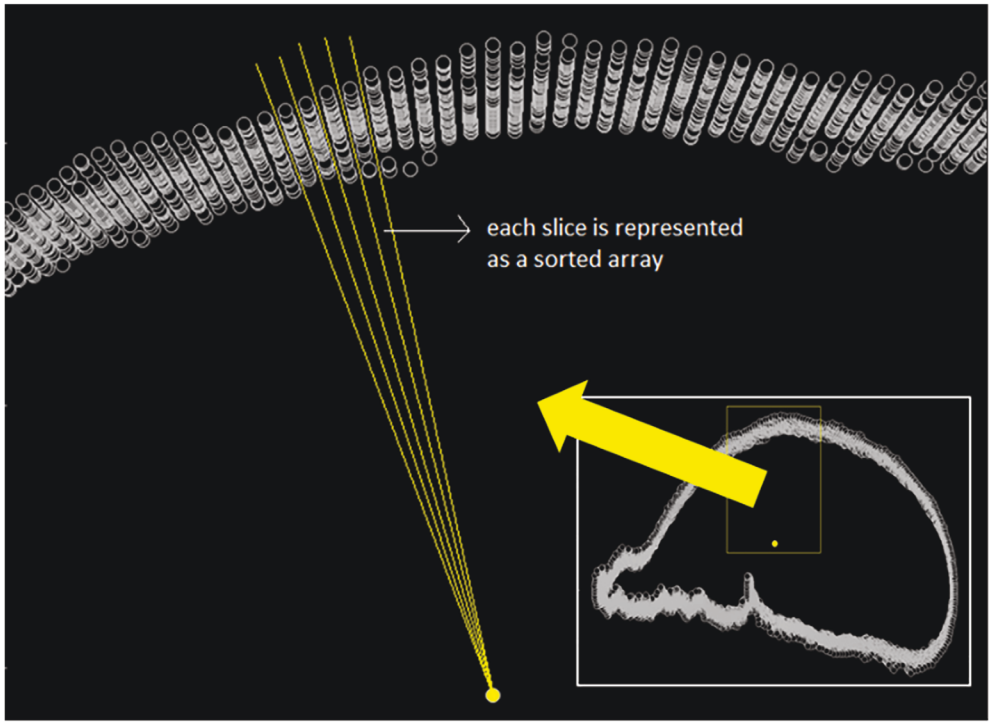
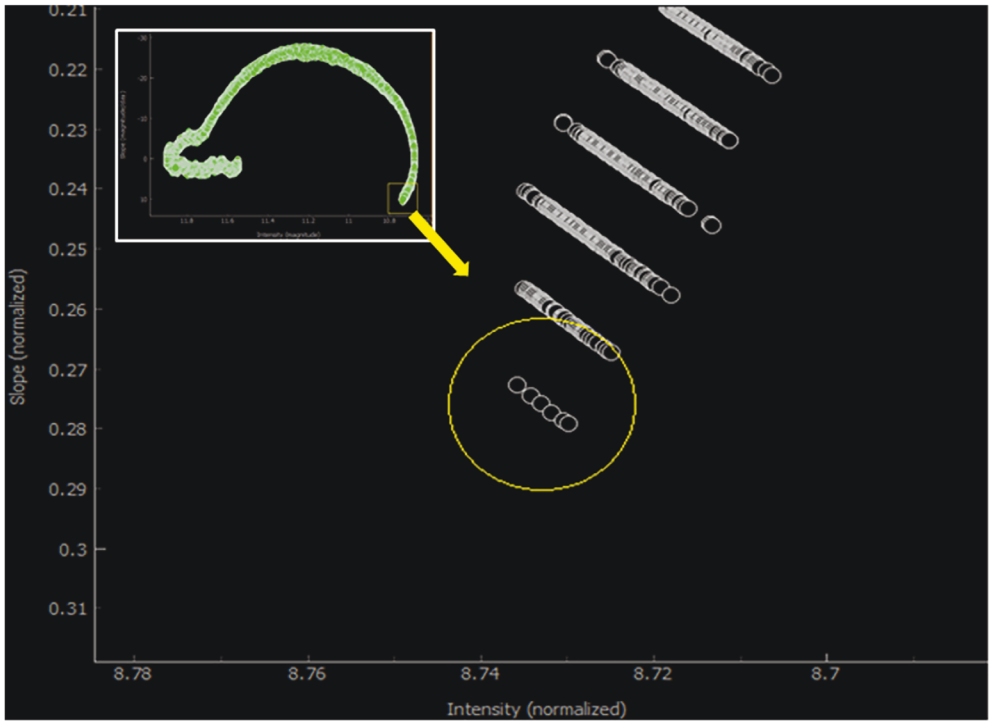
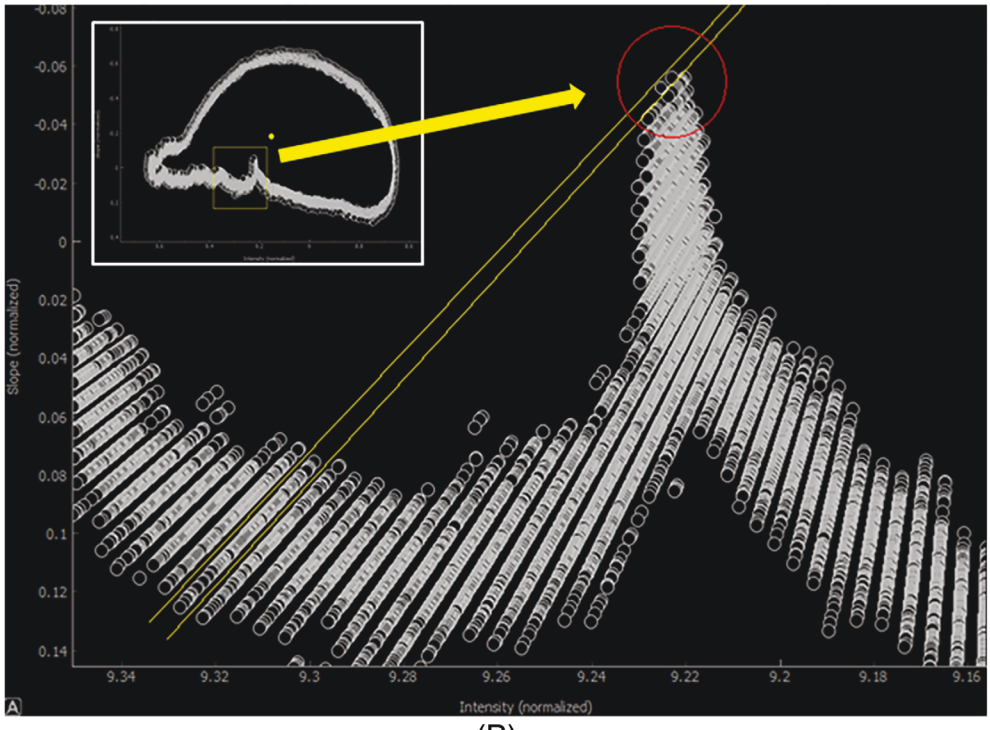


FIGURE 3 Illustration of how we internally store the 'm vs Δm' reference curve for use in real-time monitoring. The reference curve is divided into 360 slices around its midpoint. Data points that fall inside the same slice (meaning that they are within 1° of angular separation around the midpoint) are put inside a sorted one-dimensional array containing only their distances to the midpoint



(A)



(B)

FIGURE 4 Illustrations of corner cases that require special treatment. (A) Light curve with incomplete phase coverage because some parts of its oscillation always occur during the day when the telescope is not operating. The boundary region may contain only a few reference points. (B) Light curve with a region that protrudes into the middle, causing several observation points to fall into slices that contain few of their actual closest neighbors

these slices, the algorithm will simply calculate the average distance to all the available nearest neighbors in the slice. For observations that fall into empty slices, we can either choose to assign a very large deviation value or simply raise an alarm immediately. Although not currently implemented, it is also possible to use data points from neighboring slices to infer whether the newly observed data point still lies within the normal expected range.

The second problem that might occur is that for some stars, certain parts of their reference curve may protrude sharply into the middle area of the plot. When this happens, data points that lie near the edges of these protrusions may fall into slices that have very few of their actual nearest neighbors (Figure 4B). The consequence is that observations that are otherwise close to many reference data points can be misclassified as anomalies. To prevent such errors, the algorithm is set to stop looking for the next nearest neighbor when the distance to the last neighbor is more than 0.1 unit farther than the previous closest neighbor (note that the ranges in both x and y-axis are normalized to 1). As an illustration, suppose an observation falls into a slice that contains reference data points whose distances from itself, sorted in ascending order, are as follows: [0.003, 0.004, 0.15, 0.16, …]. Since the distance from the third nearest neighbor is more than 0.1 units farther than the distance to the second nearest neighbors, only the first two nearest neighbors will be used to calculate the average deviation, giving a value of 0.0035.

3.3 | Setting the dynamics threshold

After a deviation score or the dynamics distance is generated, the next decision we need to make is the dynamics threshold value at which we should raise an alarm. Here we use a similar approach to the one previously proposed for the OGLE mission.[18,19] The algorithm is set to raise an alarm whenever the deviation score exceeds 12 times the historical RMS (Root-Mean-Square) deviation (for each individual star) for 20 consecutive times. Based on our experiments, this parameter combination results in a good balance between precision and recall, while still allowing the algorithm to provide an early notification for ongoing microlensing events.

4 | DATASETS

In order to assess the performance of the DS algorithm, we tested it on two different light curve datasets. The first dataset contains typical sinusoidal light curves generated with different parameters so we can evaluate the performance of the proposed algorithm under very different scenarios. This dataset is identical to those used in the NFD algorithm[20] that has been successfully used in GWAC detection. The second dataset contains more practical light curves generated from a set of Fourier coefficients estimated from real variable stars[21] and this dataset is very close to the GWAC setting. We will inject simulated random noise and microlensing signals with varying lengths and magnitudes to the background light curves and check if the proposed algorithm can detect these microlensing signals.

4.1 | Sinusoidal light curve dataset

The first test set contains sinusoidal light curves with added noise and microlensing signals, generated with the same parameter combinations used in the "Large-scale Simulative Data Set" in the NFD study.[20] The dataset contains a total of 262,440 light curves. The background sine waves were generated for a total duration of 56 days. To simulate the real-world operating condition of GWAC, the light curves were simulated to be discontinuous, with observations occurring from 8 pm to 4 am each day, with 15 s intervals.

The microlensing signals were simulated using the following formula[22,23]:

$$u_t = \sqrt{u_0^2 + \left(\frac{t - t_0}{t_E}\right)^2}$$

$$A_t = \frac{u_t^2 + 2}{u_t\sqrt{u_t^2 + 4}}$$

$$m_t = -2.5 \log_{10} A_t$$

TABLE 1 The parameters of the added Gaussian noise and microlensing signal

Φ	σ	u_0	t_E
$\left[0, \frac{1\pi}{3}, \frac{2\pi}{3}, \frac{3\pi}{3}, \frac{4\pi}{3}, \frac{5\pi}{3}\right]$	0.01, 0.05, 0.1	0.1, 0.5, 1.0	0.005, 0.034, 0.063, 0.092, 0.121, 0.15

Note: Φ = phase of the background light curve, σ = SD of the random Gaussian noise, u_0 = microlensing parameter which controls its peak magnitude, t_E = microlensing parameter which controls its duration.

where m_t refers to the change in magnitude at time t; t_0 refers to the time when the microlensing effect is at its maximum; u_0 refers to the minimum angular separation at t_0 (controls the maximum peak of the microlensing signal), and t_E is the Einstein time, which controls how long the microlensing lasts. The signals were placed on the light curve so that their peaks were centered on day 55 ($t_0 = 55$). Microlensing signals weaker than 0.01 magnitude are clipped to 0 to prevent misclassification of real detection of weak microlensing signals as false positives.

Real-time monitoring starts from day 40, while the first 40 days of data is used to simulate past historical data. Since the microlensing signal is centered on day 55, this means that the algorithm has around a 15-day window in which it can produce a false alarm. If a false alarm is produced, monitoring will immediately stop and the light curve will be classified as a false positive. If the algorithm successfully raises an alarm during the period in which the microlensing signal is present, the light curve will be counted towards the true positive category. On the other hand, if the algorithm fails to detect any anomaly after the end of the light curve, it will be classified as a false-negative. Accuracy is defined as the number of true positives divided by the total number of the tested light curves.

In addition to accuracy, we also calculated the average alert delay of the algorithm. Since the simulated microlensing signals are symmetrical in their temporal dimension, we simply denote the start and end point of a microlensing event with 0% and 100%, respectively. An alert delay of 50% means that the algorithm raises an alarm when the event is already 50% underway, or when the microlensing effect is at its peak.

4.2 | Practical light curve dataset

To test the performance of our algorithm in a more realistic setting, we generated practical light curves using Fourier coefficients estimated from real variable stars, obtained from an online database.[21] We filtered the database for stars with periods of 2 days or less, resulting in a total hit of 1429 light curves. The phase of the background light curve (Φ) and the parameters for the added noise (σ) and microlensing signals (u_0 and t_E) are varied using the combinations shown in Table 1, for a total of 462,996 (1429*6*3*3*6) light curves. All other details including the total length of the light curves, the length, and interval of observations, where the microlensing peaks were placed on the background light curve, as well as the accuracy and alert delay calculations, were identical to the previous dataset (Section 4.1).

5 | EXPERIMENTAL RESULTS

Based on the two datasets, we measure the false positive and false-negative rate of the algorithm as well as its alert delay. We also show how various parameters can affect such results. A comparison with the existing algorithm NFD that has been employed in the GWAC observation has been given. A discussion on the time and memory requirement of the algorithm is also presented. We use Python to quickly implement our algorithm using about 3000 lines of code.

5.1 | Results from sinusoidal light curve dataset

The DS algorithm has markedly better overall accuracy on the sinusoidal light curve dataset (85.6%). Furthermore, DS on average is able to detect microlensing signals at a much earlier stage (30.0% average alert delay). Over the roughly 15 days of observations, only 0.011% of the light curves in the sinusoidal dataset produced false positives. This means that if around 176,000 stars are being monitored by each GWAC camera, we can expect to get only 1.3 false positives per night for each camera.

One of our main goals in developing the DS algorithm was to design a method that has stable performance across a variety of background light curves and microlensing signals. In order to assess how different parameters affect how our algorithm behaves, we analyzed how its accuracy and alert delay changed with different parameter combinations. The results are presented in Figure 5. Since abnormal slopes for certain magnitude values will be captured by DS easily, it performs better at recognizing anomaly events that generate larger slope deviations.

As Figure 5A shows, DS's accuracy gets progressively lower as the magnitudes of the microlensing events were lowered (their u_0 increased). While DS's alert delay will become better (shorter) with larger u_0 or t_E (see Figure 5B,D). More interestingly, the algorithm's accuracy was generally higher for microlensing events that occurred across shorter durations (t_E) (Figure 5C), since shorter durations generally cause greater slope deviations. This means that, unlike the NFD algorithm which is more sensitive to anomaly events with longer durations, DS is better suited for identifying short-timescale events.

Another factor we examined was whether the ratio between the peak magnitude of the microlensing event (denoted as A_m) and the amplitude of the background wave (A) affects DS's performance. As expected, a larger A_m/A ratio led to a higher accuracy rate (Figure 5E). Although the range in accuracy appears large, it is important to note that for small A_m/A ratios, only samples with the weakest microlensing signals (highest u_0) were included, since samples with $A > 1.5$ had been filtered out.[20] When the complete dataset before the filtration step was analyzed, even at the lowest A_m/A ratio, the algorithm still achieved around 75% accuracy.

Our results show that in general, the algorithm performed better at detecting microlensing events when the period of the background light curve was shorter, producing both slightly shorter alert delay and higher accuracy (Figure 5G,H). The likely explanation for this phenomenon is that light curves with shorter periods will naturally tend to have greater separation between their minimum and maximum slope. Therefore, this makes it easier for the algorithm to identify anomalies whose effects are relatively significant. If the separation between the minimum and maximum slope is small, then this middle area will be overfilled with reference data points, reducing our ability to distinguish anomalous signals. Also notice, however, that the algorithm's accuracy seems to improve again for the highest period we tested ($T = 1$ day). While it is difficult to draw any conclusion from only a single value, it is possible that the underlying cause of this phenomenon is that as the period gets very large relative to the window size of 15 min, the star will behave more and more like a constant star with just a single constant near-zero slope, for which anomaly detection would be easier.

Out of all the parameters we altered, the amount of noise present in the light curve seemed to be one of the most significant factors that influenced DS's performance. As with the other parameters we tested, noise level did not seem to be correlated with DS's false positive rate. Instead, it negatively affects the algorithm's sensitivity towards anomaly events. While the false-negative rate of the algorithm at 0.005 noise level was exactly 0%, it rose to 27.2% at the highest noise level of 0.2 (Figure 5I). At the same time, the algorithm's alert delay increased from 10.5% to 37.7% (Figure 5J). It is easy to see why a high noise level hinders DS's performance. As the light curve becomes noisier, the calculated m and Δm values for historical data points will have greater variance (the reference plot will become thicker). This reduces DS's ability to discriminate against anomalous signals in the 'm vs Δm' plot. It is possible to reduce this variance by lengthening the time window used to calculate the average m and Δm. However, this would also reduce the algorithm's ability in identifying anomaly events with shorter timescales. Investigating the interrelationships between window length, noise level, and the shape of the anomaly signal in an effort to determine the optimal algorithm parameters for different conditions is one potential avenue for further research.

5.2 | Results from practical light curve dataset

When tested against the dataset containing more diverse types of complex light curves, our DS algorithm was still able to maintain high accuracy (94.8%) and low average alert delay (22.5%). 0.095% of the light curves resulted in false positives. Even though this rate is higher than for the sinusoidal dataset, note that the higher overall accuracy means that there is more room for us to tone down the sensitivity of the algorithm. The algorithm's behavior as the background light curve and microlensing parameters were varied were similar in both the sinusoidal and the practical light curve dataset.

As explained before, one of the most important determinants of DS's performance is the level of noise present in the background light curve. Since the main focus of the practical dataset was to test the performance of the algorithm on a variety of light curve shapes, the number of parameter combinations for the added noise and microlensing signal was intentionally reduced to keep the total number of light curves manageable. This likely explains the higher accuracy on this dataset. Note, however, that based on our experience handling a small scale 'mini-GWAC' dataset obtained from the National Astronomical Observatories of China,[24] the maximum size of noise we used in the practical dataset

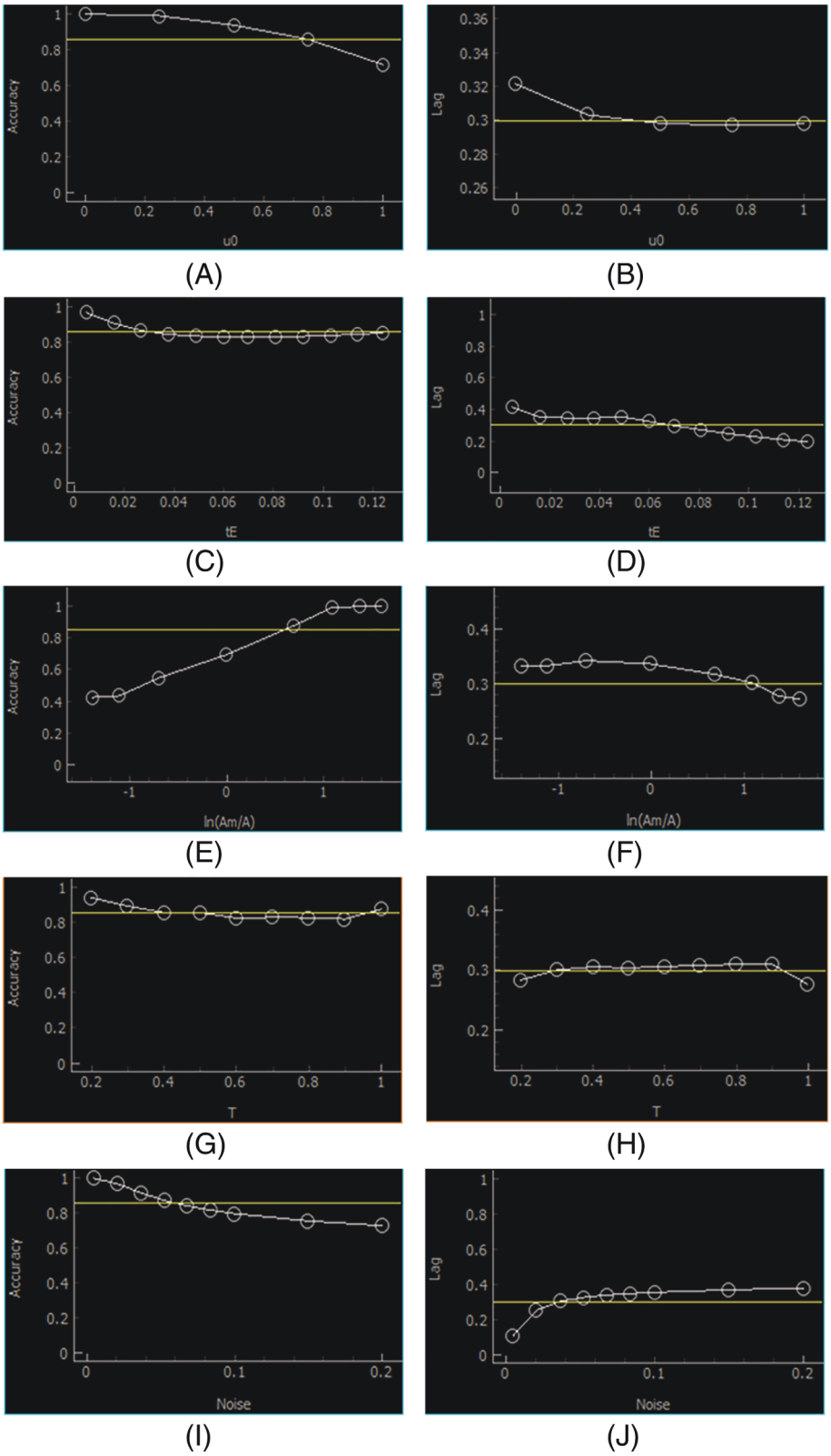


FIGURE 5 The accuracy (left column) and alert delay (right column) of DS algorithm on the sinusoidal dataset as different parameters were varied. From top to bottom: 1. u_0 2. t_E (both control the shape of the microlensing signal) 3. $\ln(A_m/A)$ (log of the ratio between the max effect size of the microlensing signal and the peak amplitude of the background sine wave) 4. T (the period of the background sine wave) 5. Noise level. Note that the scale in the y-axis may be different for each graph. The yellow horizontal lines show the average for the whole dataset

TABLE 2 Comparison with the existing algorithm used in GWAC observation

Method	Data sets	Accuracy	False positive	Lag
NFD	Sinusoidal	67.2%	X	43%
DS	Sinusoidal	85.8%	0.019%	31.2%
	Practical	94.8%	0.095%	22.5%

($\sigma = 0.1$ magnitude) is already sufficiently large. Furthermore, some of the changes we made to the microlensing parameters, that is, cutting the lower range of u_0 and increasing the upper range of t_E, should actually make it more difficult for DS to perform its detection. Overall, we believe that the results on the practical dataset should be more representative of how the algorithm will perform in real-life situations.

5.3 | Performance comparison

To evaluate the effect of the proposed DS algorithm, we compare its results with the existing NFD algorithm that has been used in the real GWAC data processing. Table 2 shows the comparison results.

NFD did not have the related experiments on the practical dataset so we do not show the result of the practical dataset for NFD. At the same time, NFD also did not provide the false positive result so we show it with an "X".

From Table 2 we can see that for the sinusoidal dataset, NFD's accuracy is 67.2% and DS can improve the accuracy to 85.8% because the combined (m, Δm) can capture the changes in both displacement and velocity of a star's brightness. Furthermore, the alert time of DS is also earlier than NFD about 11.8%.

For the practical dataset, since the light curves are built based on real stars and the noise parameter is set based on the GWAC observation data, the dataset can represent the real GWAC scenario well. The results show that both the accuracy and the alert time are better than the sinusoidal dataset. The accuracy is improved by 9% and the alert time is reduced by about 0.7%. The results show that DS algorithm is promising to achieve better performance under the practical scenarios.

5.4 | Time and memory requirement

During real-time detection, the DS algorithm needs to perform the following computations at each timestep:

- Simple averaging and a linear regression using data from the last 15 min (around 60 data points) to obtain m and Δm
- conversion of the (m, Δm) coordinate pair into its polar form and indexing into the correct slice
- locating the 20 nearest neighbors in the slice using binary search and calculate the average distance to those neighbors (simple subtractions)

On an Intel Core i7-6500U 2.5 GHz CPU, all of these steps altogether take less than 90 ms to complete for a single star. Here our large light curves processing problem is an embarrassingly parallel problem that means the light curves from different stars can be processed in parallel independently. So for more light curves, we can use more computing resources to achieve scalable performance. The total processing time required for the estimated 176,000 stars observed by a single GWAC camera can be brought down to less than 1 s with only 16 CPUs. The DS algorithm therefore far exceeds the minimum required performance needed for the GWAC system. Further code optimization would further improve the performance of the algorithm.

During maintenance time (normally daytime, but can also be done concurrently in the background), the reference curve for each star needs to be regenerated. Our testing shows that with the reference dataset size set to 30 days, the amount of time required to produce the reference curve for each star is around 0.1 s. This means that the reference plot for all 176,000 stars can be prepared in under 20 min using 16 CPUs. Notice that this refers to the amount of time needed to completely regenerate the reference plots from scratch. In practice, since it is likely that the midpoint will be relatively

stable from 1 day to the next, in most cases we can skip recalculating the polar coordinates for each data point and re-sorting all slices at every maintenance session.

The use of the heuristic method described in Section 3.2 allows us to compress historical data points from two-dimensional data to single-dimensional data, reducing the algorithm's memory requirement by almost one-half. Assuming 8 h of continuous observation per day, with a history size of 30 days, we would need to store as many as 57,600 reference data points for each star. Since each data point can be stored using a single 32-bit floating point number, the total memory requirement for 176,000 stars would be less than 48 GB, well within the normal capacity of enterprise workstations.

6 | RELATED WORK

There have been several microlensing detection projects, such as OGLE,[10] MACHO,[11] EROS,[12] KMTNet.[13] All the work has advanced the research in microlensing detection and time-domain astronomy. The data analysis problem in astronomy has become much more challenging and we must design more suitable methods to solve the problem.[25] GWAC focuses on very short-timescale (from several hours to even minutes) anomalies. At the same time, the super large field of view and high cadence observation of GWAC bring a great challenging data processing problem. In this article, our method focuses on GWAC like wide-field and high cadence survey data.

Rebbapragada et al.[26] proposed an unsupervised algorithm PCAD to identify microlensing events from catalogs of periodic variable stars. This phase alignment method can help to handle unsynchronized periodic time-series data. Rajpaul[27] presented an evolutionary algorithm to perform autonomous fitting of gravitational microlensing lightcurves. Price-Whelan et al.[28] presented a statistical search method to recover microlensing events in uniformly sampled data and this work can help constrain all-sky event rate predictions and tests microlensing signal recovery in large data sets. Nun et al.[29] introduced a novel mixture of the experts' outlier detection model to achieve better results than any single expert model using three fields from the MACHO catalog and generated a list of anomalous candidates. Chen et al.[30] employed a hierarchical Gaussian process to create a general and stable model of time series for anomaly detection, and apply this approach to the light curve problem. Godines et al.[31] used a random forest algorithm to search for microlensing in wide-field surveys even with low-cadence data. All these methods target offline survey data with relatively long timescales, the proposed method focuses on large online survey data with very short timescales.

The earliest attempt for a real-time detection system of microlensing signals which occur on variable stars was described by Wyrzykowski et al.[23] Bozza et al.[32] proposed a method to analyze weak microlensing events in real-time. The use of machine learning methods such as the hierarchical temporal memory (HTM)[33] and the long short term memory (LSTM) Neural Networks[34] have been proposed as potential ways to perform anomaly detection on time series data. However, our experiments show that such approaches are commonly too memory- and time-consuming for use in real-time settings involving large-scale data. Feng et al.[24] explored the use of a modified ARIMA model called DARIMA (Dynamic Auto-Regressive Integrated Moving Average) which can dynamically adjust its model parameters during real-time processing of time series data. Kim et al.[13] proposed an algorithm focusing on detecting "rising" events. Qiu et al.[20] proposed a simple and easy method to implement an online NFD algorithm to detect outliers from millions of stars. Our method is inspired by all the related real-time methods and has better performance for detecting weak microlensing events.

Several important systems have been developed to detect microlensing, such as the real time data analysis systems in the OGLE-III survey,[19] the Korea Microlensing Telescope Network (KMTNet) Alert Algorithm, and Alert System.[13] A software package MulensModel[35] was also developed for gravitational microlensing modeling and this kind of software will be very helpful for analyzing the microlensing events efficiently.

Beyond astronomy detection, anomaly detection has been widely employed in many different fields.[1] Hardware Trojan detection[36] can be solved using information entropy. Clustering the information entropy of different circuit logics can detect hardware Trojans. Anomaly detection can also solve some decision-decision problems. The uncertainty measure method using fuzzy soft sets[37] can help to make decisions and an application has been used in COVID-19 diagnosis. Relative outlier distance and biseries correlations methods in[38] are used to distinguish the anomaly from changing points that have significant changes but are in a normal state. In popular wireless sensor networks, Safaei et al.[39] provides an anomaly detection method with limited computational and communication capabilities, and nine time series related features are used in their method. Online multi-source VMware monitoring[40] is an application very close to our scenario and a semi-supervised technique where a flat adaptive clustering technique has been used to build the training model.

We build our detection in phase space based on domain knowledge so our detection effect is much better than the general clustering method.

This article provides a novel dynamics view to detect the anomaly in multi-dimensional dynamics feature space. The advantage of the proposed method is that it can identify weak anomalies that are hard to be detected by other methods. At the same time, this view will also be helpful to analyze the dynamics behaviors of detected anomalies.

7 | CONCLUSION

This article provides a novel view to investigate the anomaly of a periodic variable star in brightness space with typical dynamics metrics. The concept of DS in a multi-dimensional dynamics features space is proposed to identify the anomalous changes from the baseline behaviors. The DS can be regarded as the normal track so any changing state far away from the track should be an anomaly. This is the basic idea of our anomaly detection method.

We employ two typical dynamics metrics (displacement and velocity) to capture the motion of a star in brightness space and construct a star's DS in two-dimensional feature space. Our method can detect the anomaly caused either by displacement or by velocity or both of them, so it is much sensitive to detect weak anomalies that cannot be identified by the brightness only criteria.

Based on the basic idea of DS, we develop a highly optimized DS algorithm to identify anomalies from millions of stars' online survey data in real-time. Taking advantage of the feature of a periodic variable star, we calculate K (in this article we select $K = 20$) nearest neighbors to measure the deviation degree from given star's Dynamic Signature or the 'm vs Δm' reference curve in our implementation. Polar coordinates are used to save half memory for signature data and locate the nearest neighbors quickly. The high performance of the proposed DS algorithm makes it easy to handle the online big data and trigger alarms at a very early stage of a microlensing event. Our linear regression method can generate stable brightness (displacement) and brightness changing speed (velocity) values, and this is very helpful to improve the accuracy of our algorithm.

Extensive experiments are conducted using two different kinds of datasets. The Sinusoidal light curve dataset is used to investigate how our DS algorithm's performance will be affected by the microlensing parameters (u_0, t_E), magnitude ratio (A_m/A) between the microlensing event (A_m) and the background wave (A), the period of a given variable star and the noise level in the survey data. The experimental results show that our algorithm can achieve both higher accuracy and shorter delay between the start of the microlensing event and the time of detection under a wide range of parameter space. We also generated practical light curves using Fourier coefficients estimated from real variable stars to evaluate our algorithm under a practical scenario. Different phases of the background light curve, different SDs of the random Gaussian noise and different microlensing parameters (u_0 and t_E) are used to evaluate our method and experimental results show that our DS algorithm can gain even better performance both in accuracy (94.8%) and delay (22.5%) than that on the simulated dataset.

Our algorithm provides a powerful method for allowing one to easily identify anomalous data. This can be done even in situations where the time series has many discontinuities, and when the period or the mathematical model of the time series cannot be easily estimated beforehand. Although the current study focuses on gravitational microlensing, the same principle can potentially be applied in other domains which involve anomaly detection in periodic time series.

ACKNOWLEDGMENTS

This research is supported in part by the Key Research and Development Program of China (No. 2016YFB1000602), the "Key Laboratory of Space Astronomy and Technology, National Astronomical Observatories, Chinese Academy of Sciences, Beijing, 100012, China", the National Natural Science Foundation of China (No. 61440057, 61272087, 61073008, 61762074 and 11690023), and the MOE Research Center for Online Education Foundation (No. 2016ZD302). This research was funded in part by NSF Grant number CCF-2109988 (Bader).

AUTHOR CONTRIBUTIONS

Ivan Hendy Goenawan: implement the experiment and design the draft method. provide the draft version. **Zhihui:** propose the basic idea and algorithm design and improve the proposed method; finish the final version. **Chao Wu:** propose the idea on domain related algorithm design and the evaluation method; contribute the algorithm design. **Yankui Sun:**provide the comparison method. **Jianyan Wei:** provide the domain related algorithm design method. **David A. Bader:** provide the algorithm optimization design method.

DATA AVAILABILITY STATEMENT

The datasets used in our article are available in two ways. (1) For the synthetic data, anyone can generate them based on the parameters given in the article. (2) for the more practical light curves, the related Fourier coefficients are available from https://urldefense.com/v3/__https://nitro9.earth.uni.edu/fourier/__;!!N11eV2iwtfs!79859wztbDvnNKEdlm7K8KsftM4qpshFr0p-nSizCTqZrtfTYfxIFevbGmVXmfbX$

ORCID

Zhihui Du https://orcid.org/0000-0002-8435-1611
Yankui Sun https://orcid.org/0000-0001-7155-8261

REFERENCES

1. Chandola V, Banerjee A, Kumar V. Anomaly detection: a survey. *ACM Comput Surv*. 2009;41(3):15-58.
2. Moniez M. Microlensing as a probe of the galactic structure: 20 years of microlensing optical depth studies. *Gen Relativ Gravit*. 2010;42(9):2047-2074.
3. Beaulieu JP, Bennett DP, Fouqué P, et al. Discovery of a cool planet of 5.5 earth masses through gravitational microlensing. *Nature*. 2006;439(7075):437-440.
4. Motch C, SVOM Consortium with the help from Cordier B, Lachaud C, Jianyan W, and many other colleagues. The SVOM mission. *Astron Nachr*. 2017;338(9–10):978-983.
5. Sackett PD. Planet detection via microlensing; 1997. Arxiv preprint astro-ph/9709269.
6. Burgdorf MJ, Bramich DM, Dominik M, et al. Exoplanet detection via microlensing with RoboNet-1.0. *Planet Space Sci*. 2007;55(5):582-588.
7. McGhee, M. (2018). *Detecting Primordial Black Hole Microlensing Events*. https://dc.uwm.edu/uwsurca/2018/Posters/101/
8. Gaudi BS. Microlensing surveys for exoplanets. *Annu Rev Astron Astrophys*. 2012;50:411-453.
9. Mao S. Astrophysical applications of gravitational microlensing. *Res Astron Astrophys*. 2012;12(8):947-972.
10. Tsapras Y, Hundertmark M, Wyrzykowski Ł, et al. The OGLE-III planet detection efficiency from six years of microlensing observations (2003–2008). *Mon Not R Astron Soc*. 2016;457(2):1320-1331.
11. Alcock C, Allsman RA, Alves DR, et al. The MACHO project: microlensing optical depth toward the galactic bulge from difference image analysis. *Astrophys J*. 2000;541(2):734-766.
12. Afonso C. Discovery and photometry of the binary-lensing caustic-crossing event EROS-BLG-2000-5; 2003. arXiv preprint astro-ph/0303647.
13. Kim HW, Hwang KH, Shvartzvald Y, et al. The Korea microlensing telescope network (KMTNet) alert algorithm and alert system; 2018. arXiv preprint arXiv:180607545.
14. MOA Group. The microlensing observations in astrophysics. *Nature*. 2011;473:349-352.
15. Cordier B, Wei J, Atteia JL. The SVOM gamma-ray burst mission; 2015. arXiv preprint arXiv:151203323.
16. Wei J, Cordier B, Antier S, et al. The deep and transient universe in the SVOM era: new challenges and opportunities-scientific prospects of the SVOM mission; 2016. arXiv preprint arXiv:161006892.
17. Heinze AN, Tonry JL, Denneau L, et al. A first catalog of variable stars measured by the asteroid terrestrial-impact last alert system (ATLAS). *Astron J*. 2018;156(5):241.
18. Udalski A, Szymanski M, Kaluzny J, et al. The optical gravitational lensing experiment. The early warning system; 1994. arXiv preprint astro-ph/9408026.
19. Udalski A. The optical gravitational lensing experiment. Real time data analysis systems in the OGLE-III survey; 2004. arXiv preprint astro-ph/0401123.
20. Qiu J, Sun Y, Wu C, Du Z, Wei J. NFD: toward real-time mining of short-timescale gravitational microlensing events. *Publ Astron Soc Pac*. 2018;130(992):104504.
21. Morgan SM. WWW database of variable star Fourier coefficients. *Publ Astron Soc Pac*. 2003;115(812):1250-1255.
22. Albrow M, Beaulieu JP, Birch P, et al. The 1995 pilot campaign of PLANET: searching for microlensing anomalies through precise, rapid, round-the-clock monitoring. *Astrophys J*. 1998;509(2):687-702.
23. Wyrzykowski Ł, Rynkiewicz AE, Skowron J, et al. OGLE-III microlensing events and the structure of the galactic bulge. *Astrophys J Suppl Ser*. 2015;216(1):12.
24. Feng T, Du Z, Sun Y, Wei J, Bi J, Liu J. Real-time anomaly detection of short-time-scale GWAC survey light curves. Proceedings of the 2017 IEEE International Congress on Big Data (BigData Congress); 2017:224-231; IEEE.
25. Huijse P, Estevez PA, Protopapas P, Principe JC, Zegers P. Computational intelligence challenges and applications on large-scale astronomical time series databases. *IEEE Comput Intell Mag*. 2014;9(3):27-39.
26. Rebbapragada U, Protopapas P, Brodley CE, Alcock C. Anomaly detection in catalogs of periodic variable stars. *Astronomical Data Analysis Software and Systems XVIII*. Vol 411; Astronomical Society of the Pacific (ASP); 2009:264.
27. Rajpaul V. *A Novel Algorithm for Analysing Gravitational Microlensing Events*. Doctoral dissertation. University of Cape Town; 2012.
28. Price-Whelan AM, Agüeros MA, Fournier AP, et al. Statistical searches for microlensing events in large, non-uniformly sampled time-domain surveys: a test using palomar transient factory data. *Astrophys J*. 2014;781(1):35.

29. Nun I, Protopapas P, Sim B, Chen W. Ensemble learning method for outlier detection and its application to astronomical light curves. *Astron J*. 2016;152(3):71.
30. Chen H, Diethe T, Twomey N, Flach PA. Anomaly detection in star light curves using hierarchical Gaussian processes. ESANN; 2018.
31. Godines D, Bachelet E, Narayan G, Street RA. A machine learning classifier for microlensing in wide-field surveys. *Astron Comput*. 2019;28:100298.
32. Bozza V, Dominik M, Rattenbury NJ, et al. OGLE-2008-BLG-510: first automated real-time detection of a weak microlensing anomaly-brown dwarf or stellar binary? *Mon Not R Astron Soc*. 2012;424(2):902-918.
33. Ahmad S, Purdy S. Real-time anomaly detection for streaming analytics; 2016. arXiv preprint arXiv:160702480.
34. Malhotra P, Vig L, Shroff G, Agarwal P. Long short term memory networks for anomaly detection in time series. Proceedings of the 23rd European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN); 2015:89; Presses universitaires de Louvain.
35. Poleski R, Yee JC. Modeling microlensing events with MulensModel. *Astron Comput*. 2019;26:35-49.
36. Lu R, Shen H, Feng Z, Li H, Zhao W, Li X. HTDet: a clustering method using information entropy for hardware Trojan detection. *Tsinghua Sci Technol*. 2021;26(1):48-61.
37. Bhardwaj N, Sharma P. An advanced uncertainty measure using fuzzy soft sets: application to decision-making problems. *Big Data Mining Anal*. 2021;4(2):94-103.
38. Ji C, Zou X, Liu S, Pan L. ADARC: an anomaly detection algorithm based on relative outlier distance and biseries correlation. *Softw Pract Exper*. 2019;50:1-17. doi:10.1002/spe.2756
39. Safaei M, Ismail AS, Chizari H, et al. Standalone noise and anomaly detection in wireless sensor networks: a novel time-series and adaptive Bayesian-network-based approach. *Softw Pract Exper*. 2020;50:428-446. doi:10.1002/spe.2785
40. Solaimani M, Iftekhar M, Khan L, Thuraisingham B, Ingram J, Seker SE. Online anomaly detection for multi-source VMware using a distributed streaming framework. *Softw Pract Exper*. 2016;46(11):1479-1497.

How to cite this article: Goenawan IH, Du Z, Wu C, Sun Y, Wei J, Bader DA. Dynamics signature based anomaly detection. *Softw: Pract Exper*. 2021;1-16. doi: 10.1002/spe.3052