Publications of the Astronomical Society of the Pacific, 134:024201 (20pp), 2022 February https://doi.org/10.1088/1538-3873/ac44f0

Citizen ASAS-SN Data Release. I. Variable Star Classification Using Citizen Science

C. T. Christy[1], T. Jayasinghe[1,2], K. Z. Stanek[1,2], C. S. Kochanek[1,2], Z. Way[3], J. L. Prieto[4], B. J. Shappee[5], T. W.-S. Holoien[6,9], T. A. Thompson[1,2,7], and A. Schneider[8]

[1] Department of Astronomy, The Ohio State University, 140 West 18th Avenue, Columbus, OH 43210, USA; christy.125@osu.edu
[2] Center for Cosmology and Astroparticle Physics, The Ohio State University, 191 W. Woodruff Avenue, Columbus, OH 43210, USA
[3] Department of Physics and Astronomy, Georgia State University, Atlanta GA 30303, USA
[4] Núcleo de Astronomía de la Facultad de Ingeniería y Ciencias, Universidad Diego Portales, Av. Ejército 441, Santiago, Chile
[5] Institute for Astronomy, University of Hawai'i, 2680 Woodlawn Drive, Honolulu, HI 96822, USA
[6] The Observatories of the Carnegie Institution for Science, 813 Santa Barbara Street, Pasadena, CA 91101, USA
[7] Millennium Institute of Astrophysics, Santiago, Chile
[8] ASC Technology Services, 433 Mendenhall Laboratory 125 South Oval Mall Columbus OH, 43210, USA

Received 2021 December 6; accepted 2021 December 17; published 2022 February 25

Abstract

We present the first results from Citizen ASAS-SN, a citizen science project for the All-Sky Automated Survey for Supernovae (ASAS-SN) hosted on the Zooniverse platform. Citizen ASAS-SN utilizes the newer, deeper, higher cadence ASAS-SN g-band data and tasks volunteers to classify periodic variable star candidates based on their phased light curves. We started from 40,640 new variable candidates from an input list of ~7.4 million stars with $\delta < -60^\circ$ and the volunteers identified 10,420 new discoveries which they classified as 4234 pulsating variables, 3132 rotational variables, 2923 eclipsing binaries, and 131 variables flagged as Unknown. They classified known variable stars with an accuracy of 89% for pulsating variables, 81% for eclipsing binaries, and 49% for rotational variables. We examine user performance, agreement between users, and compare the citizen science classifications with our machine learning classifier updated for the g-band light curves. In general, user activity correlates with higher classification accuracy and higher user agreement. We used the user's "Junk" classifications to develop an effective machine learning classifier to separate real from false variables, and there is a clear path for using this "Junk" training set to significantly improve our primary machine learning classifier. We also illustrate the value of Citizen ASAS-SN for identifying unusual variables with several examples.

Unified Astronomy Thesaurus concepts: Variable stars (1761); Eclipsing binary stars (444); Stellar rotation (1629); Light curves (918); Stellar classification (1589); Catalogs (205); Surveys (1671)

1. Introduction

Variable stars are some of the most useful astrophysical tools as they are used to probe many aspects of stellar evolution and galactic structure. Eclipsing binaries allow the derivation of empirical calibrations for fundamental stellar parameters such as mass and radii (Torres et al. 2009). The period–luminosity relation of Cepheids is crucial to probing cosmological distances (Leavitt 1908; Freedman et al. 2019; Riess et al. 2018). The short period δ Scuti variables allow us to study the scaling relations between stellar parameters (effective temperature, surface gravity, density, etc.) and astroseismology (Hasanzadeh et al. 2021). For researchers to truly utilize these systems, it is important that they be discovered and classified.

The search for new variable stars is now dominated by large surveys. This includes surveys such as the All-Sky Automated Survey (ASAS; Pojmanski 2002), the All-Sky Automated Survey for SuperNovae (ASAS-SN; Shappee et al. 2014; Kochanek et al. 2017; Jayasinghe et al. 2018, 2021), the Asteroid Terrestrial-impact Last Alert System (ATLAS; Heinze et al. 2018; Tonry et al. 2018), the Catalina Real-Time Transient Survey (CRTS; Drake et al. 2009), EROS (Derue et al. 2002), Gaia (Prusti et al. 2016; Brown et al. 2018), MACHO Alcock et al. 2000, the Northern Sky Variability Survey (NSVS; Woźniak et al. 2004), the Optical Gravitational Lensing Experiment (OGLE; Udalski 2004), and the Zwicky Transient Facility (ZTF; Bellm 2014).

ASAS-SN is a wide-field photometric survey that monitors the entire night sky using 20 telescopes located in both the Northern and Southern hemispheres (Shappee et al. 2014; Kochanek et al. 2017; Jayasinghe et al. 2018). ASAS-SN detects variables and other transients in the process of finding bright supernovae (Holoien et al. 2016). For the initial V-band catalog of variables, ~60 million stars were classified through machine learning techniques, resulting in a catalog of

[9] NHFP Einstein Fellow.

∼426,000 variables, of which ∼220,000 were new discoveries (Jayasinghe et al. 2020, 2021).

Using machine learning techniques to identify and classify variable stars is particularly efficient for common variable classes and other known phenomena. However, some object classes are ambiguous and noise or systematic errors will sometimes confuse the classifiers. We can address this problem by using citizen science to classify variable star candidates in ASAS-SN along with machine learning. Citizen science may also more effectively identify rare phenomena compared to a machine learning classifier due to their scarcity in the training data (Alhammady & Ramamohanarao 2004). The ASAS-SN citizen science project, Citizen ASAS-SN, is hosted on the Zooniverse[10] platform and aims to assist in the classification of variable stars. The Zooniverse is the worlds largest hub for citizen science; in recent years, it has hosted many successful projects that often lead to serendipitous discoveries (Trouille et al. 2019).

Here we analyze the first results of Citizen ASAS-SN. We examine the classifications made by the citizen scientists and their ability to correctly label variable stars from their light curves. Through their classifications, we have discovered 10,420 new variable stars and flagged many interesting variables for follow-up studies. We find that citizen scientists can reliably separate "junk" sources from real variable stars and distinguish between pulsating variables and eclipsing binaries. We also outline our new g-band machine learning classifier and discuss its performance compared to the citizen scientists. In Section 2 we describe the ASAS-SN data used to generate light curves. Section 3 discusses the new g-band machine learning classifier. We outline the details of Citizen ASAS-SN in Section 4, along with an analysis of the classifications made by the citizen scientists. In Section 5 we compare the machine learning and citizen science classifications. We highlight some of the interesting variable stars our users encountered in Section 6 and discuss the utility of citizen science in identifying such systems. We present a summary of our work in Section 7.

2. The ASAS-SN g-band Catalog of Variable Stars

Starting in 2014, ASAS-SN began surveying the sky in the V-band with a limiting magnitude of $V \lesssim 17$ mag and a ∼2–3 day cadence using 8 telescopes on two mounts in Chile and Hawaii. Each ASAS-SN camera takes 3 images with 90 second exposures for each epoch. The field of view of an ASAS-SN camera is 4.5 deg^2, the pixel scale is $8\farcs0$ and the FWHM is typically ∼2 pixels. ASAS-SN uses image subtraction (Alard & Lupton 1998; Alard 2000) for the detection of transients and variable sources. Since 2018, ASAS-SN has shifted to the g-band and expanded to 20 cameras on 5 mounts, adding new units in South Africa, Texas, and Chile. All of the ASAS-SN telescopes are hosted by the Las Cumbres Observatory (LCO; Brown et al. 2013). When compared to the V-band data, the g-band data has an improved depth ($g \lesssim 18.5$ mag), cadence ($\lesssim$24 hr in the g-band versus ∼2–3 days in the V-band), and reduced diurnal aliasing due to the longitudinal spread of the ASAS-SN units.

As our input source catalog for this project, we used the `refcat2` catalog (Tonry et al. 2018). For this paper, we selected all sources with declinations $\delta < -60°$, $g < 18$ mag and $r_1 < 30\farcs0$, where the `refcat2` metric `r1` is the radius at which the cumulative G flux in the aperture exceeds the flux of the source being considered and is a measure of blending around a star. After applying these selection criteria, we were left with ∼7.4 million sources. We extracted their g-band light curves as described in Jayasinghe et al. (2018) using image subtraction (Alard & Lupton 1998; Alard 2000) and aperture photometry on the subtracted images with a 2 pixel radius aperture. We corrected the zero-point offsets between the different cameras as described in Jayasinghe et al. (2018) and calculated periodograms using the Generalized Lomb-Scargle (GLS, Zechmeister & Kürster 2009; Scargle 1982) algorithm.

Candidate variable sources were identified using various cuts in light curve (for e.g., median magnitude, root-mean-square deviation, and string length statistics) and GLS periodogram statistics (power and false alarm probability) as summarized in Jayasinghe et al. (2019). The Citizen ASAS-SN workflow presently focuses on the classification of periodic variable stars, so we did not include non-periodic sources in this work. We required that the false alarm probability for the period is better than 10^{-7} for sources with median magnitudes fainter than $g = 16.5$ mag. Figure 1 shows the distribution of variables by their average g-band magnitude. Variable sources with magnitudes of $g \leqslant 11.5$ were considered to be saturated. Note the peak at the faint end of the distribution. It comes from removing some candidate selection criteria used by Jayasinghe et al. (2018) with consequences we did not fully appreciate at the time (see Section 4.4).

The final product of this paper is the first installment of the ASAS-SN g-band catalog of variable stars. This includes classification data from our updated machine learning classifier as well as the input from our citizen scientists. We also include supplementary data from crossmatches to existing photometric catalogs. The revised catalog is available at https://asas-sn.osu.edu/variables.

3. g-Band Machine Learning Classifier

The machine learning classifier used for ASAS-SN's V-band variable catalogs is extensively described in Jayasinghe et al. (2019a). It was based on a `scikit-learn` (Pedregosa et al. 2018) random forest model that was trained to distinguish between broad variable types using features which included light curve statistics, Gaia distances, and multi-band

[10] Zooniverse:https://www.zooniverse.org/.

photometry. While this classifier was extremely accurate at identifying common variable types, it often mislabeled rare phenomena and light curves with systematic errors.

We retrained the random forest classifier described in Jayasinghe et al. (2019a) using features from our new *g*-band data. The training set for this updated classifier is the same as that used previously. We included two additional features based on the Lafler–Kinmann (Lafler & Kinman 1965; Clarke 2002) string length statistic (LKSL). We calculated the LKSL statistic $T(t)$ on the temporal light curve using the definition

$$T(t) = \frac{\sum_{i=1}^{\mathrm{N}}(m_{i+1} - m_i)^2}{\sum_{i=1}^{\mathrm{N}}(m_i - \overline{m})^2} \times \frac{(N - 1)}{2N} \quad (1)$$

from Clarke (2002), where the m_i are the time ordered magnitudes and $\overline{m}$ is the mean magnitude. We also calculated the LKSL statistic sorting the light curve based on phase for both the best GLS period and twice the best GLS period, which we will call $T(\phi|P)$ and $T(\phi|2P)$ respectively. For the two new classification features, we used the difference in the Lafler–Kinmann string length statistics ordered in phase using the best period and ordered as time,

$$\delta(t, P) = \frac{T(\phi|P) - T(t)}{T(t)}, \quad (2)$$

and the difference in the statistics phased by the period and twice the period,

$$\delta(P, 2P) = \frac{T(\phi|2P) - (T(\phi|P)}{T(\phi|P)}. \quad (3)$$

The ML classification pipeline automatically corrects the period as described in Jayasinghe et al. (2019b). The updated RF classifier classifies sources into 7 broad classes (CEPH, DSCT, ECL, LPV, RRAB, RRc/RRd, and ROT) which are subsequently refined into sub-classes (see Jayasinghe et al. 2019a). The overall precision, recall and F_1 parameters for the updated RF classifier are 94.4%, 95.3% and 94.7% respectively. An important feature of the ML classifier to keep in mind is that it only provides a probability for the type of variable. There is no equivalent of the ”Junk” class available to the citizen scientists in large part because there was no training set to define it.

4. Project Description and Results

Volunteers working on Citizen ASAS-SN are shown images with light curves phased by both the best GLS period and twice the best GLS period along with the observed light curve. Our goal was to address several simple but common problems distinguishing between variable types (such as RRc RR Lyrae variables and EW eclipsing binaries). For eclipsing binaries, the best period returned by the GLS periodogram is often 1/2 of the orbital period, which is why the light curve phased with twice the best GLS period is shown. When phased with the correct orbital period, the light curves of eclipsing binaries show a distinct separation of the primary and secondary eclipses allowing for their accurate classification; this behavior is shown in the example light curve shown in Figure 2. The observed light curve is useful for identifying long-period variables such as Miras, and evolving variables like rotating spotted stars.

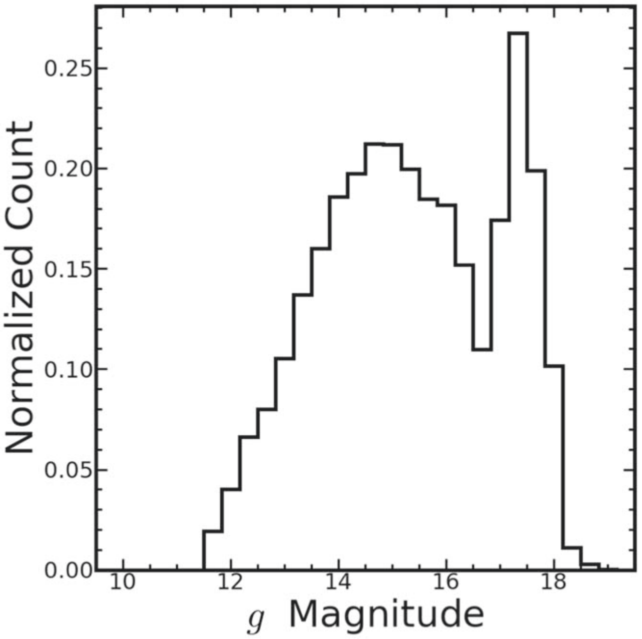


Figure 1. Number distribution of the mean *g*-band magnitudes for the candidate variable sources with $\delta < -60°$. Retrospectively, the peak at faint magnitudes mainly consists of false positive candidates located near field edges (see Section 4.4). In Jayasinghe et al. (2018) these were removed by several candidate selection criteria that were relaxed when we selected variable candidates for this project.

We designed our project workflow to be easy to navigate and accessible to a wide array of volunteers. Because we expect no prior knowledge of variable star classification, we first present users with a tutorial that details the classification process and summarizes the science. Volunteers also have access to a field guide that describes common variable stars and their light curves. Our workflow tasks users to determine the correct basic classification, selecting between three broad classes (Pulsating Variables, Eclipsing Binaries, Rotational Variables), choosing the option “Unknown Variable” for ambiguous cases, or flagging the light curve as “Junk”. Figure 2 shows an example of the workflow. As users get started, we present them with a variety of “gold-standard” (GS) candidates that have been classified by the science team. These GS variables provide the user with feedback on their classifications to train them in the process. As users make more classifications, GS variables

Table 1
Breakdown of the Number of Most Voted Classifications for each Variable Type, Including those found the AAVSO VSX (Watson et al. 2006), OGLE III (Poleski et al. 2012), and OGLE IV (Kozlowski et al. 2013) Catalogs

	All Candidates	Eclipsing Binaries	Pulsating Variables	Rotational Variables	Unknown	Junk
N Candidates	40,640	12,292	11,621	4529	161	12,037
In VSX	16,750	9018	6230	1320	29	153
In OGLE III	1132	214	819	43	1	55
In OGLE IV	2560	464	1989	75	0	32
Average Probability	0.72	0.77	0.72	0.53	0.40	0.75
New	10420	2923	4234	3132	131	0

Note. The variables listed as unknown are those with nonspecific classifications such as MISC or VAR.

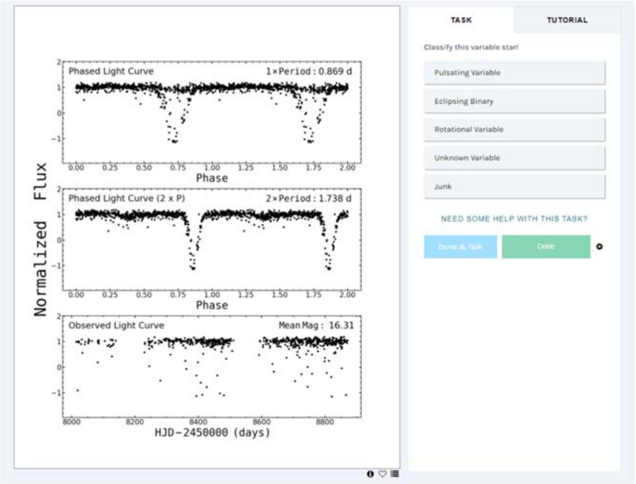


Figure 2. Citizen ASAS-SN workflow for classifying periodic variables. (Left) 2 phased light curves and the observed light curve. (Right) Possible classifications for users to select. This variable would best be classified as an Eclipsing Binary.

become less frequent and the user begins to classify new light curves.

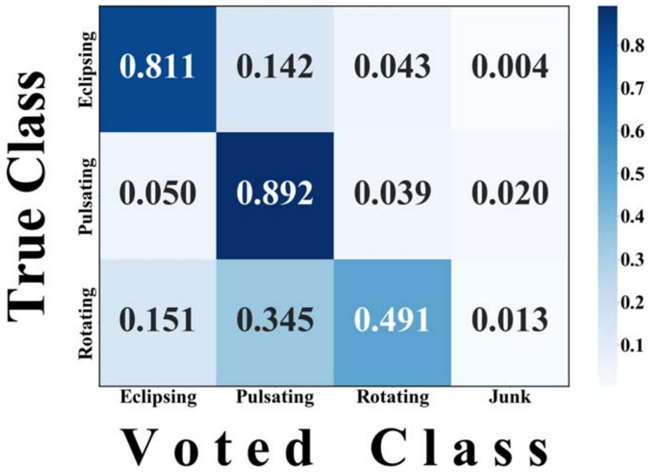
We released Citizen ASAS-SN for public use on 2021 January 5th, and it has since accrued over 3000 volunteers and ~800,000 classifications. We launched the project with a set of 40,640 variable candidates around the South celestial pole ($\delta < -60°$). In addition to the basic classification, our users pointed out many interesting variables on the project's Talk forum. We designed the workflow so that a variable candidate stops being shown to users once it has reached a retirement limit of 10 votes. If the number of "Junk" votes reaches 5, then the candidate is retired early. Once every subject was retired, we tallied up the number of votes each candidate received in each category. We then assigned each candidate a "most voted label" which describes the most popular variable type chosen by our volunteers. If a tie occurred for the most popular vote, the most voted class was chosen randomly between the tied options.

A breakdown of the most voted class for each retired variable is shown in Table 1. Of the three main variability classes (Pulsating Variables, Rotational Variables, and Eclipsing Variables), the Rotational Variable class was voted the least common, making up only 12% of the total. Pulsating variables, eclipsing binaries, and junk variable classifications were 28%, 29%, and 30%, with less than 1% classified as Unknown.

4.1. Cross-matches to External Catalogs

We cross-matched our initial subject set of 40,640 candidates with previously classified variables stars in the AAVSO VSX (Watson et al. 2006), OGLE III (Poleski et al. 2012), and OGLE IV (Kozlowski et al. 2013) catalogs using a matching radius of 16″ and found 16,750 matches in VSX, 1132 matches in OGLE III, and 2560 matches in OGLE IV. The VSX catalog contains all the variables previously identified by ASAS-SN (Jayasinghe et al. 2020). After excluding the Junk classifications, our volunteers discovered 10,420 new variables. Known eclipsing binaries made up the majority of the matches with VSX, while pulsating variables appear to be the most common match in the OGLE catalogs. A breakdown of the number of candidates accounted for by VSX, OGLE III, and OGLE IV candidates is shown in Table 1 along with the full candidate set. Table 1 also gives the average probability defined as the ratio between the number of votes for the most popular classification and the total number of votes which we will refer to as the classification strength.

Our volunteers were able to recover 99%, 95%, and 99% of the previously cataloged VSX, OGLE III, and OGLE IV variables respectively. If we define the classifications in these catalogs as the "true class" and the the most popular Citizen ASAS-SN classification as the "voted class", we find the confusion matrix shown in Figure 3. Our users could reliably distinguish between the three broad variability types, with pulsating variables as the most identifiable. Overall we found that our users correctly identified 81% of known eclipsing binaries, 89% of known pulsating variables, and 49% of known rotational variables. The poor performance on the rotational



True Class \ Voted Class	Eclipsing	Pulsating	Rotating	Junk
Eclipsing	0.811	0.142	0.043	0.004
Pulsating	0.050	0.892	0.039	0.020
Rotating	0.151	0.345	0.491	0.013

Figure 3. The normalized confusion matrix between the citizen science voted classifications on the horizontal axis and the true class classification based on the AAVSO VSX, OGLE III, and OGLE IV catalogs.

variables was expected because the morphology of their light curves can vary widely, leading to inconsistent classifications (e.g., Thiemann et al. 2021).

4.2. User Performance

For our first set of candidates, a total of 2298 volunteers participated in Citizen ASAS-SN and they made 403,626 classifications. Of these, 370,277 were of the variable candidates and 33,349 were of the gold-standard variables. We found that 1594 users made classifications from accounts registered with the Zooniverse platform, while 704 users made classifications from unregistered accounts. The registered users contributed to 95% of the total classifications, while unregistered users contributed 5% of classifications.

The next metric we considered was how correlated our user's votes were with each other for each variable type. To do this, we computed a "Classification Strength" P as the ratio between the number of votes for the voted classification type and the total number of votes. This metric would be $P = 1.0$ if all user classifications agree for a particular variable candidate. There is a lower bound of $P = 0.2$, where the 10 votes were evenly divided over the five possible classifications. In Figure 4, we show the distribution of classification strengths for each variable class. The mean classification strength was highest for eclipsing, pulsating, and junk variables with averages of $\langle P \rangle = 0.78$, 0.73 and 0.74 respectively. For candidates most voted as Rotational Variable and Unknown Variable, there are more disagreements between users with mean classification strengths of $\langle P \rangle = 0.53$ and 0.37 respectively. The low classification strength for rotating variables is in agreement with the poor performance shown in Figure 3. Given the nature of unknown variable types, a low classification strength is to be

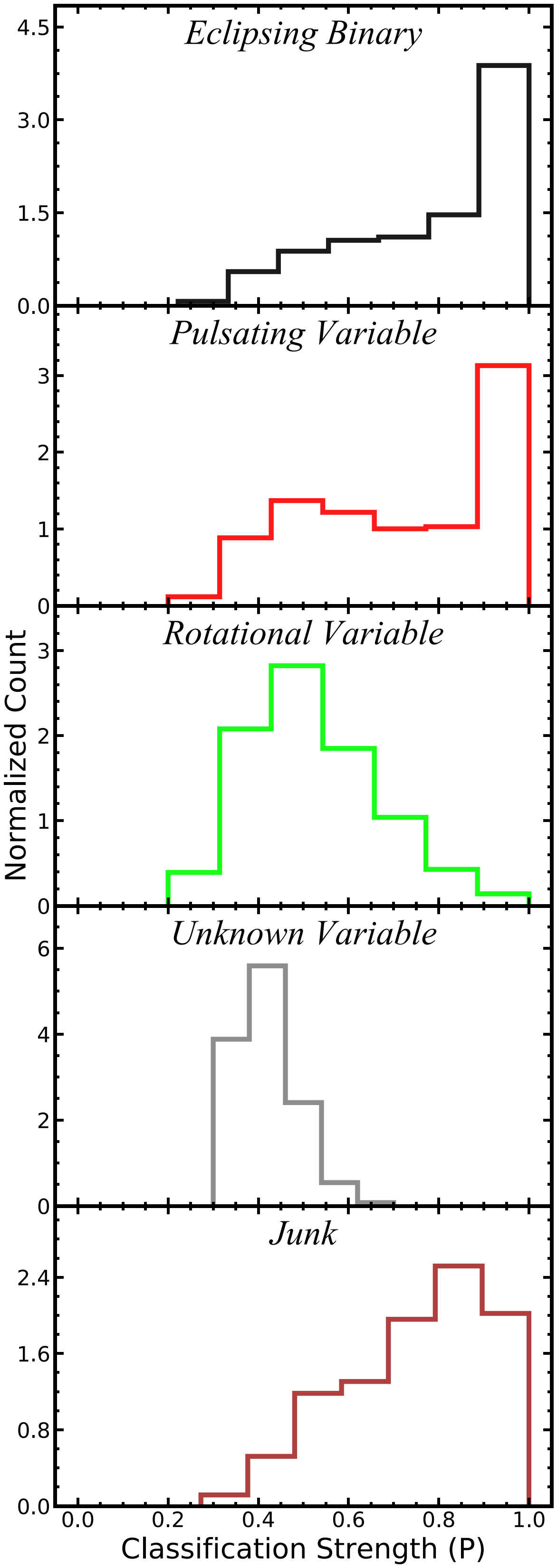
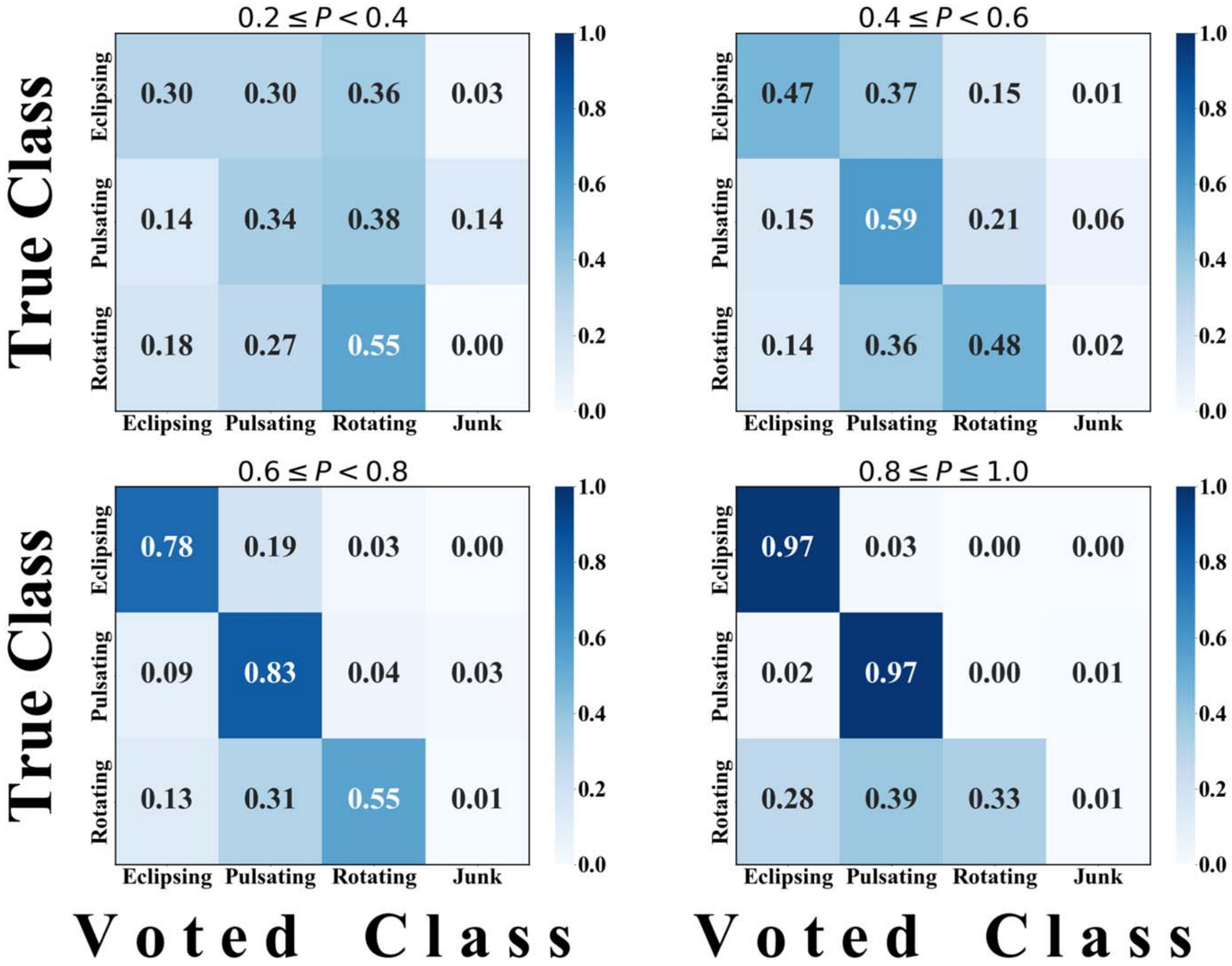


Figure 4. Normalized distribution of classification strengths for each variable class.



$0.2 \le P < 0.4$

True \ Voted	Eclipsing	Pulsating	Rotating	Junk
Eclipsing	0.30	0.30	0.36	0.03
Pulsating	0.14	0.34	0.38	0.14
Rotating	0.18	0.27	0.55	0.00

$0.4 \le P < 0.6$

True \ Voted	Eclipsing	Pulsating	Rotating	Junk
Eclipsing	0.47	0.37	0.15	0.01
Pulsating	0.15	0.59	0.21	0.06
Rotating	0.14	0.36	0.48	0.02

$0.6 \le P < 0.8$

True \ Voted	Eclipsing	Pulsating	Rotating	Junk
Eclipsing	0.78	0.19	0.03	0.00
Pulsating	0.09	0.83	0.04	0.03
Rotating	0.13	0.31	0.55	0.01

$0.8 \le P \le 1.0$

True \ Voted	Eclipsing	Pulsating	Rotating	Junk
Eclipsing	0.97	0.03	0.00	0.00
Pulsating	0.02	0.97	0.00	0.01
Rotating	0.28	0.39	0.33	0.01

Figure 5. Confusion matrices for known variable classes against voted classes binned by classification strength (P).

expected, as the class was designed to encapsulate difficult to classify variables and anomalous light curves.

In Figure 5, we show the confusion matrix (see Figure 3) for 4 ranges of classification strength. As the classification strengths increase, the performance of the citizen scientists improves for the eclipsing, pulsating and junk categories. But for rotational variables, a higher classification strength does not translate into better performance. In fact, their performance was worst in the higher classification strength bin, although this could be a statistical fluke because few rotational variables had such high classification strengths.

Light curves for variable candidates with high classification probabilities ($P = 1.0$) are shown in Figures 6, 7, and 8. Our users all agreed on their classifications of these candidates and found them easy to classify. We show examples of candidates that our users had difficulty classifying (i.e., with low classification probabilities, $P < 0.5$) in Figure 9. Sources with low classification probabilities typically displayed atypical pulsation patterns or were near our detection limits.

4.3. Grading

The average user of Citizen ASAS-SN made ∼17 classifications, of which ∼3 were for Gold Standard (GS) targets and ∼14 were for our new candidates. The distribution of the total number of new candidate classifications made by each volunteer is shown in Figure 10. We graded each user based on the number of times they agreed with the most popular classification as a proxy for the correct classification. Users with very few classification submissions produce the peaks at 0.0, 0.5, and 1.0. We divided our users into active and inactive groups, where an active user was one who submitted more than the median number of non-GS classifications ($N_{\rm Class} > 14$). The

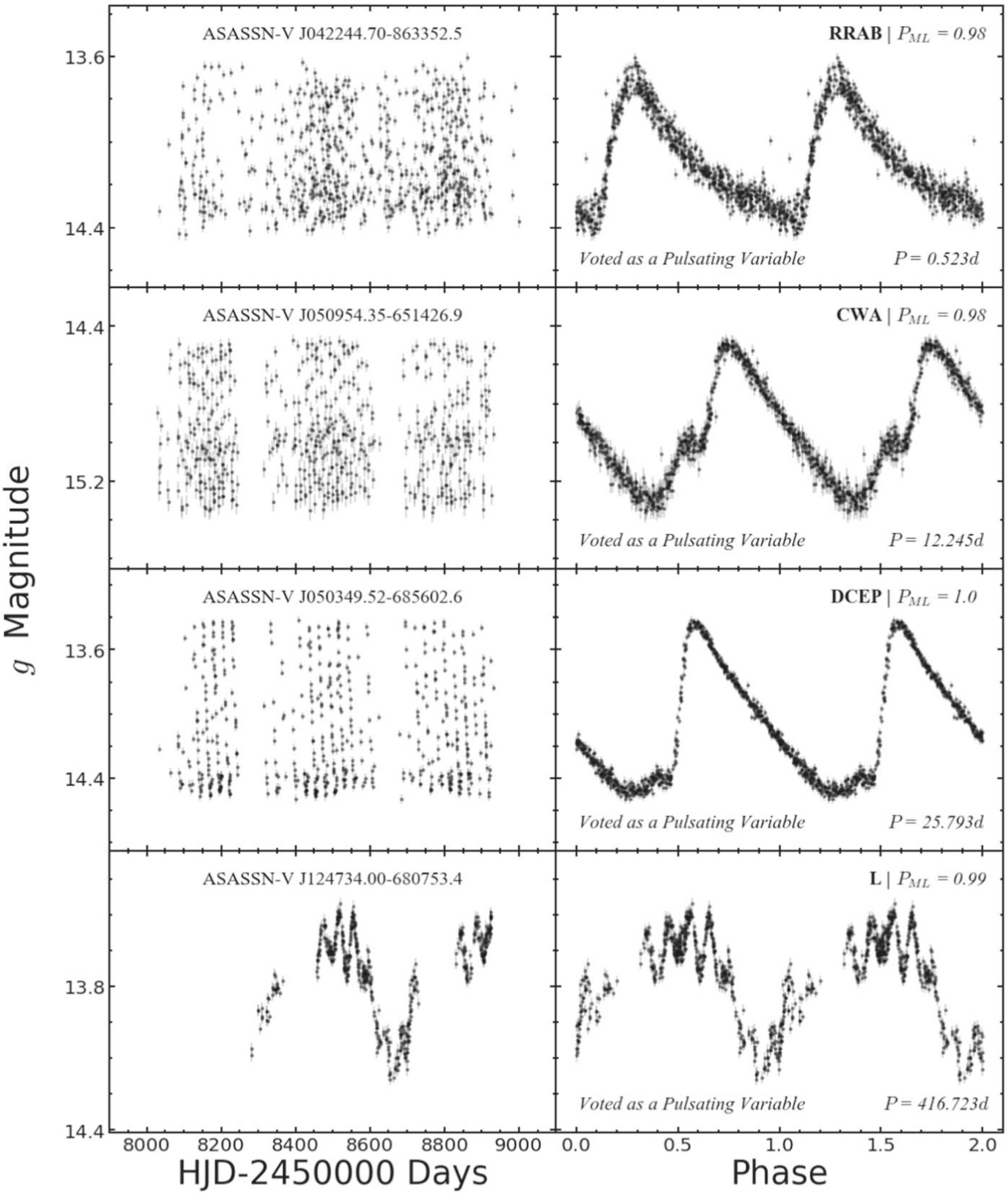


Figure 6. Light curves for a random sample of pulsating variables, with classification probabilities of 1.0, meaning all classifiers agreed on the variable type. The machine learning classification and its probability P_{ML} are given in the upper right corner.

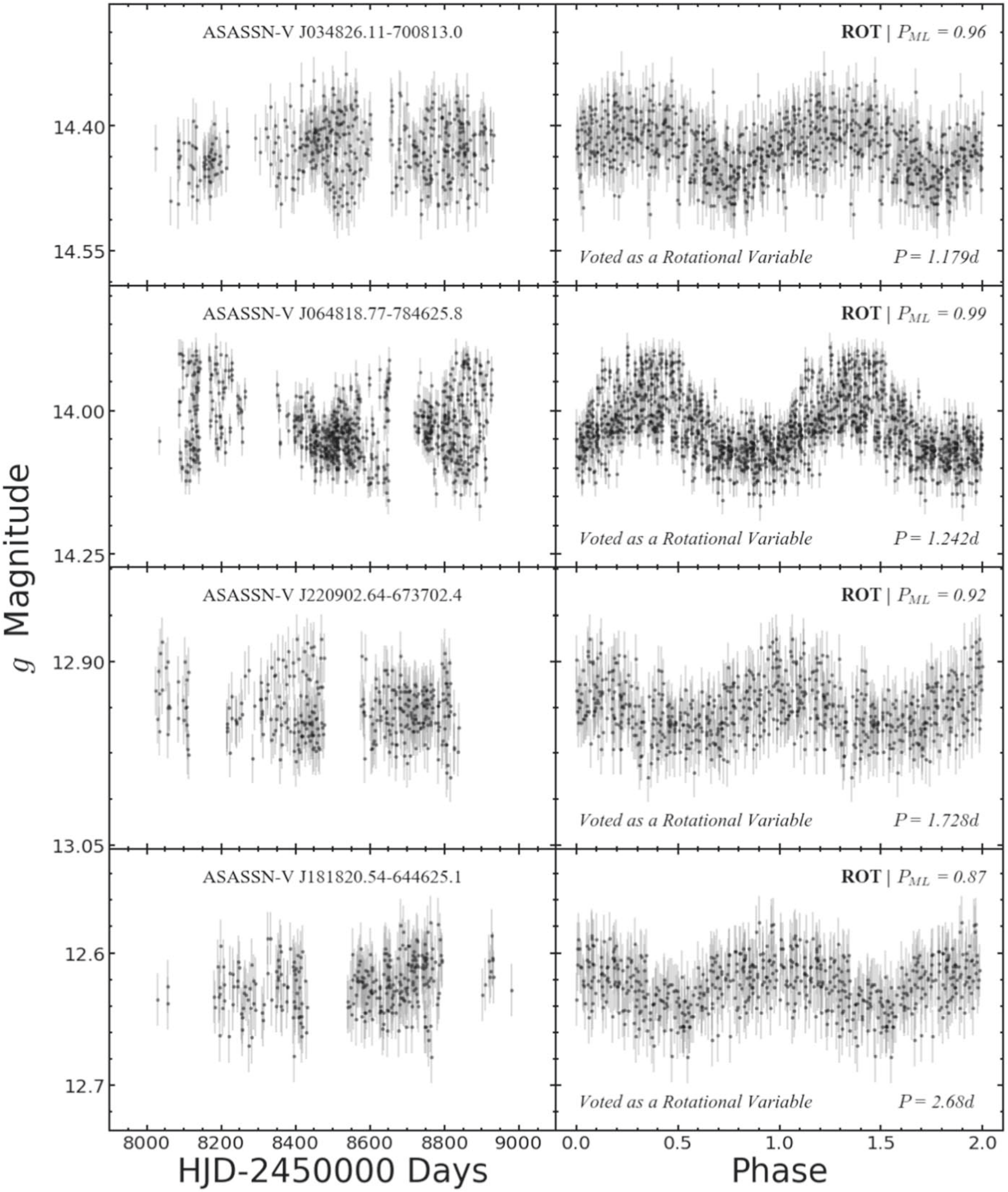


Figure 7. Light curves for a random sample of rotational variables, with classification probabilities of 1.0, meaning all classifiers agreed on the variable type. The machine learning classification and its probability $P_{\rm ML}$ are given in the upper right corner.

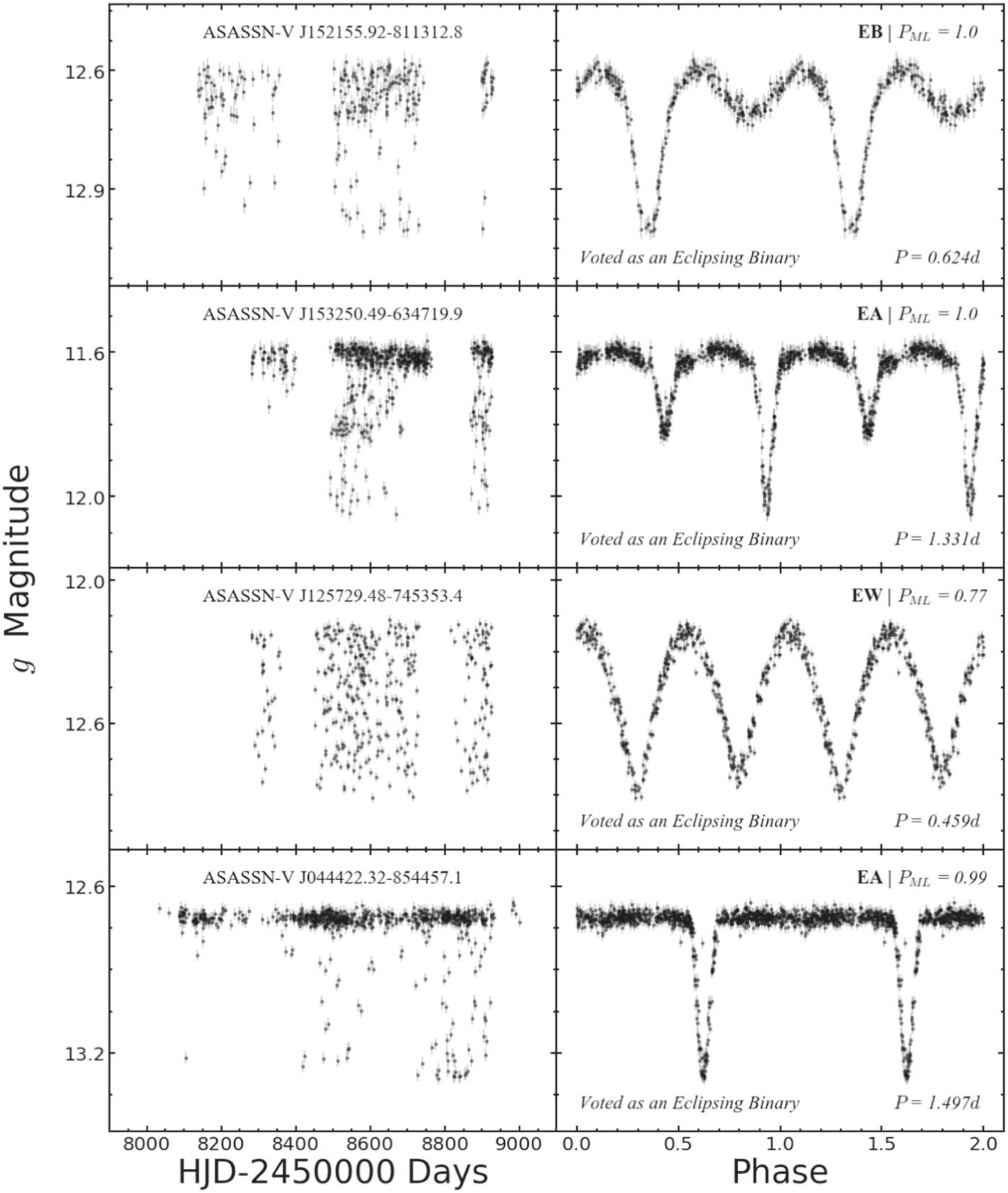


Figure 8. Light curves for a random sample of eclipsing binaries, with classification probabilities of 1.0, meaning all classifiers agreed on the variable type. The machine learning classification and its probability P_{ML} are given in the upper right corner.

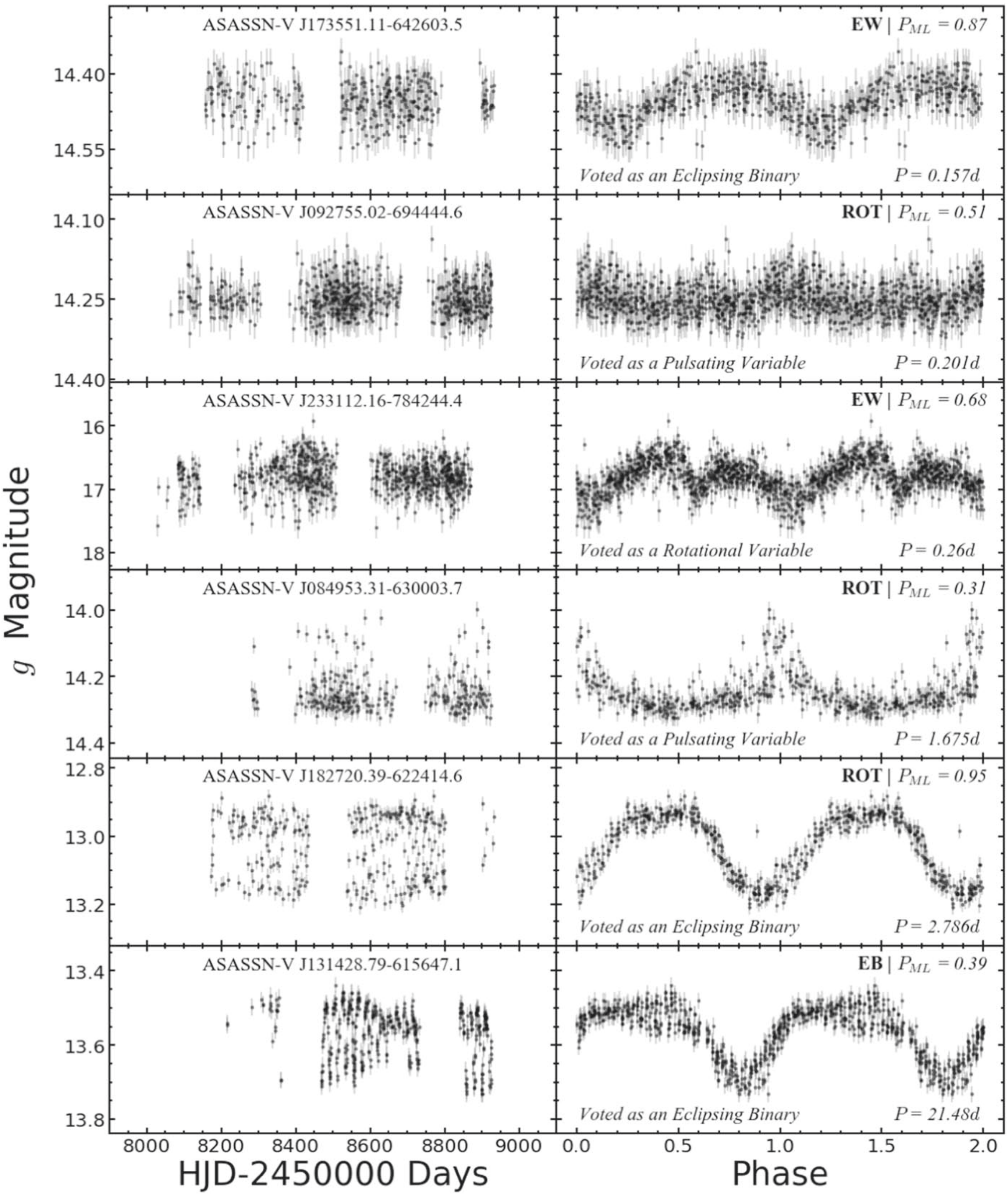


Figure 9. Light curves for candidates with low probability classifications; classification probability < 0.5. Users found these light curves difficult to classify. The machine learning classification and its probability $P_{\rm ML}$ are given in the upper right corner.

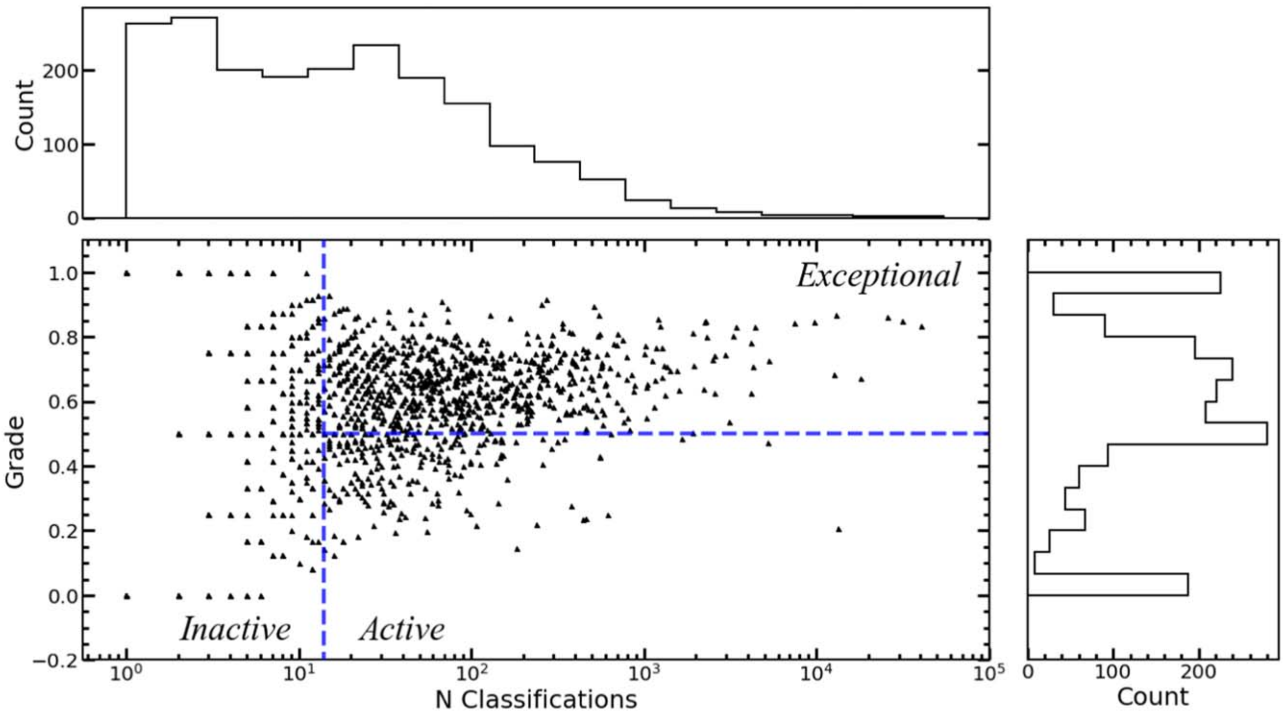


Figure 10. Distribution of the users in their number of classifications and their grade defined by the fraction of time they joined the majority vote. Histograms show the projected distribution of each quantity. The dashed lines show the division of the users into the inactive ($N_{\rm Class} \leqslant 14$), active ($N_{\rm Class} > 14$), and exceptional ($N_{\rm Class} > 14$ and grade >0.5) groups.

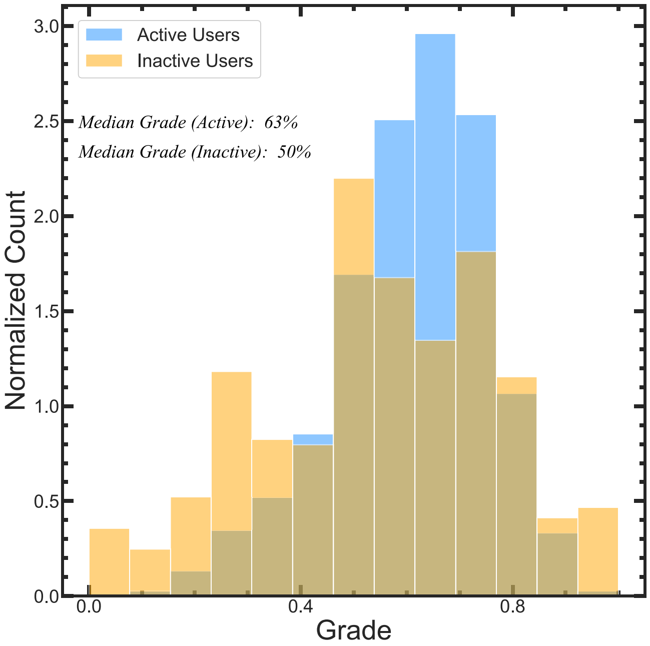


Figure 11. Normalized grade distributions for active users and inactive users. Active users are defined as users who made more than the median number of classifications ($N > 14$). We excluded inactive users with less than 3 classifications to lessen the peaks at 0.0, 0.5, and 1.0.

Table 2
Breakdown of user Grades and Candidate Classification Counts for all, Active ($N_{\rm Class} > 14$), and Inactive Users ($N_{\rm Class} \leqslant 14$).

	$N_{\rm Users}$	$\tilde{\rm Grade}$	$\tilde{N}_{\rm Class}$	Total $N_{\rm Class}$
All Users	1982	0.60	14	370277
Active Users	975	0.63	59	365652
Inactive Users	1007	0.50	3	4625

distribution of the grades and total classification count for the users is shown in Figure 10.

While 2298 users classified objects, only 1,982 classified some of the new candidates and the remaining only looked at GS targets. We only assigned grades to users who classified non-GS light curves. Of these, 975 were active and 1007 were inactive. Although inactive users were the majority, they contributed a negligible number of classifications. Of the 370,277 candidate classifications, 365,652 (99%) were made by active users. The median number of classifications made by active users was 59 while inactive users had a median of 3. Table 2 summarizes the user performance for all, active, and inactive users. As shown in Figure 11, the active members of the project outperformed the inactive group in terms of voter agreement, with active users receiving a median grade of 63% compared to 50% for inactive users.

In Figure 10, the grades of the active users appear to increase with the number of classifications and there is less scatter. There also appears to be a lower bound to this distribution proportional to log ($N_{\rm Class}$). We defined the 766 active users ($N_{\rm Class} > 14$) with grades greater than 0.5 as exceptional. Figure 13 compares the confusion matrices for the exceptional users to the inactive ($N_{\rm Class} \leqslant 14$) or poor active ($N_{\rm Class} > 14$ and grade $\leqslant$0.5) users. The highly graded active users were much better at correctly classifying known variable stars compared to inactive and low scoring users. The better performance is presumably a combination of learning, interest, and motivation.

4.4. New Discoveries

After fully classifying these new ASAS-SN variable candidates in the southern sky, our users have helped us discover over 10,000 new variable sources that are not present in the existing VSX, OGLE III and OGLE IV variable star catalogs. A breakdown of the number of new variables and their most voted variable types are shown in Table 1.

When narrowing down our initial 40,640 candidates sources, we first removed all candidates that were voted as ”Junk” by our users. Figure 12 shows the sky distribution of our full candidate set and the confirmed variables. The full candidate set displays concentric rings of artifacts associated with the lower signal to noise field edges created by the vignetting of the telescopes. Retrospectively, we also found that they mostly had periods of ∼1 day or ∼1 lunar month and g magnitudes near our detection limits (see Figure 1). In the V-band catalog (Jayasinghe et al. 2020), these were being automatically rejected because sources very close to these periods were not considered as candidates, but we had dropped this restriction when selecting candidates for this study. We found that the candidates producing the concentric ring pattern were systematically classified as Junk, and there are no patterns in the sky maps once the Junk candidates are removed. This shows that citizen science is an effective tool for cleaning data sets of false positives. We have, however, added selection criteria so that these false positives are now automatically removed (see Section 5).

After removing the Junk candidates, we cross-matched our sample with existing variable star catalogs to identify the known variables. This resulted in a sample of 10,420 new ASAS-SN discoveries. The positions of these new variable stars on the sky are also shown in Figure 12, along with the candidate, non-junk, and cross-matched sets. Of the new variable sources, the biggest subset was pulsating variables with 4234 found by our users. Rotational variables were the next most common with 3132 sources, and eclipsing binaries were the least common with 2923 sources. Our users also classified 131 of the new variables sources as unknown variables with difficult to classify light curves.

5. Machine Learning and Citizen Science

Using the updated RF classifier, we classified the candidate set and and separated the outputs into Junk and non-Junk groups. We compare the machine learning classifications of the ∼28,000 non-Junk candidates and the variables with known classifications in Figure 14. The comparison between the machine learning classifications and the most voted class by our users shows the same pattern as in Figure 3. The g-band RF classifier agreed with our users' classifications 77%, 90%, and 55% of the time for eclipsing, pulsating, and rotating variables respectively. Figure 14 also shows a confusion matrix comparing the machine learning classifications to the classifications for known variables. Here the agreement is much stronger at 96%, 93%, and 82% for eclipsing, pulsating, and rotating variables respectively. Compared to our citizen scientists, the g-band classifier was much more efficient at classifying known variable stars in our candidate sample. We recognize that this is a bit circular as some of the known variables were either used to train the ML classifier or classified by the ASAS-SN V-band machine learning classifier. Using the more refined classifications from the g-band classifier, we show the M_G versus $G_{BP} - G_{RP}$ color–magnitude diagram and the M_G versus $G_{BP} - G_{RP}$ period–luminosity diagram for the non-junk candidates broken down by type in Figure 15. The positions for each subclass of variables agrees with the distribution of variable stars in the ASAS-SN V-band catalog Jayasinghe et al. (2019a).

The ML classifier assigns a probability for each light curve to be a particular type of variable, and we adopt the highest probability classification and the frequencies of these classifications for the Junk and non-Junk sources are shown in Figure 16. The type distributions of the Junk and non-Junk sources are quite different, with many of the Junk sources placed in the non-specific VAR class. As shown in Figure 17 the ML classification probabilities for the Junk and non-Junk stars are also very different—the classification probabilities of the non-Junk sources are strongly peaked near unity with a median $P_{\rm best} = 0.95$, while the Junk sources had a median of $P_{\rm best} = 0.51$. We investigated the Junk sources with high ML probabilities and generally agreed with the citizen scientists, although there were some real but low amplitude variables. The treatment of the Junk sources by the ML classifier illustrates a standard shortcoming of machine learning. As trained, it has to classify every light curve as a variable, but for Junk sources it ”compensates” by having low classification probabilities and by putting most of them into the least well-defined variable type (generic VAR).

One approach to a solution would be to simply use the mismatched distribution in classification probability to try to automate the elimination of the Junk sources. Figure 17 (upper

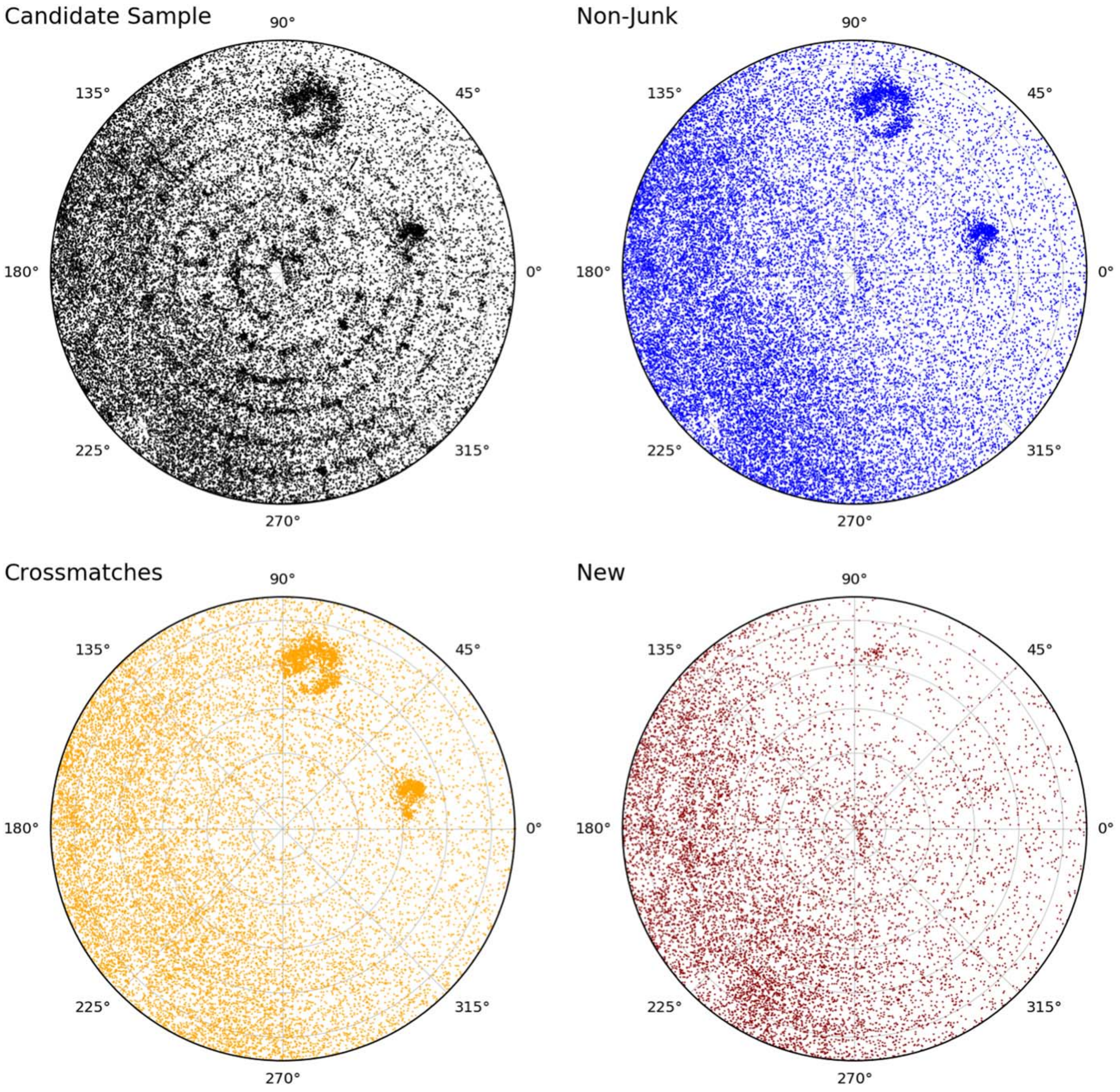


Figure 12. Radial projection of variable candidates around the South equatorial pole (top, left), variable candidates not voted as Junk (top, right), cross-matches to known VSX, OGLE III, and OGLE IV variables (bottom, left), and new variables (bottom, right). The concentric rings seen in the top left panel are due to spurious variables along field edges.

right panel) also shows the fraction of non-Junk sources as a function of the classification probability. The variable sample can be made very pure, but such a sample would also be quite incomplete. For example, if we simply keep things with classification probabilities greater than the median probability for the Junk sources, we lose 10% of the real variables while still have half of the junk sources.

A better ML solution is to use the availability of Junk and non-Junk training sets to train a new random forest classifier to distinguish them. We split the DR1 sample and used 40% of it for training and 60% for testing. The resulting classifier had an F1 score of 95.4% and precision/recall scores for non-Junk and Junk sources of 98%/92% and 96%/96%, respectively. The bottom panel of Figure 17 shows the distribution of the sources

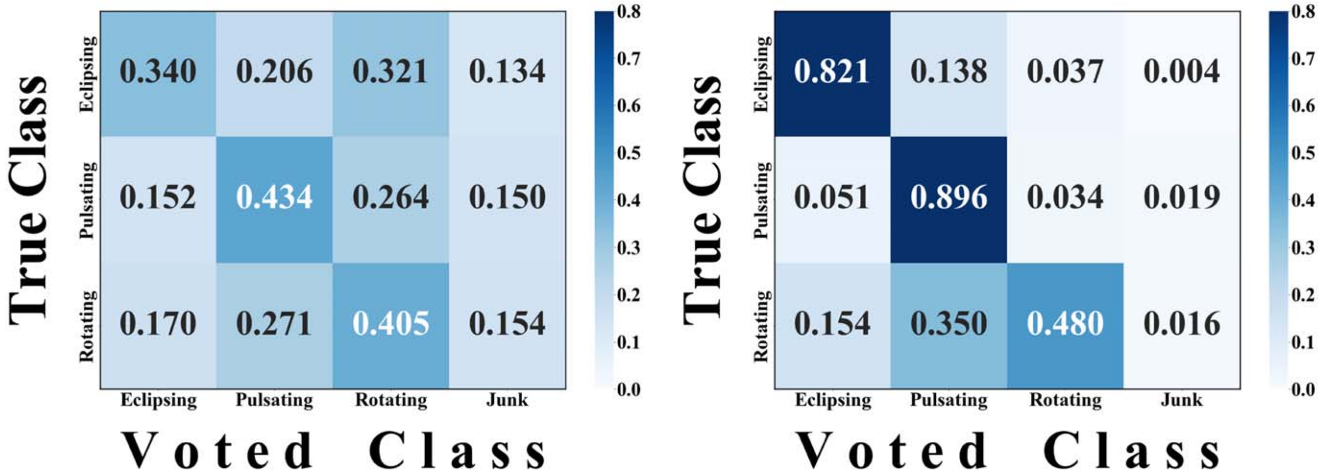


Figure 13. (Left) Confusion matrix for inactive and low graded users; users with a grade $\leqslant 0.5$ or $N_{\text{Class}} \leqslant 14$ classifications. (Right) Confusion matrix for high graded active users; users with a grade >0.5 and $N_{\text{Class}} > 14$ classifications.

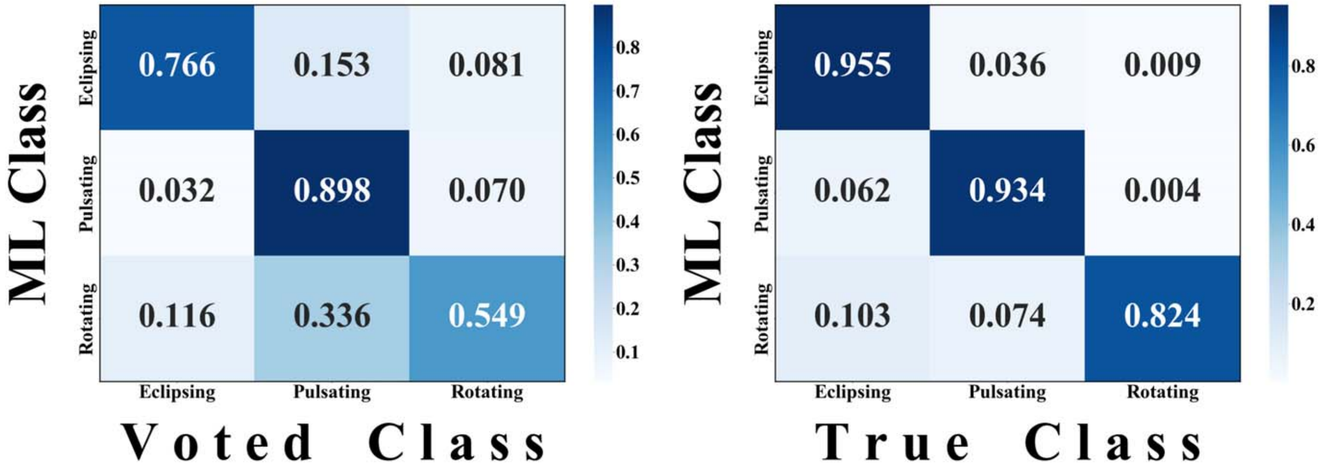


Figure 14. (Left) Confusion matrix for machine learning classifications against the most voted class. (Right) for machine learning classifications against the classification of known variable stars.

in the Junk classification probability. Keeping only sources with a less than 50% Junk classification probability eliminates roughly 97% of the Junk while losing only 3% of the variables. The upper right panel of Figure 17 shows the sample purity as function of the original variable classification probability for various cuts on the Junk probability. Clearly the path forward is to fully incorporate a Junk class into the g-band ML classifier.

6. Unusual Variables

When users encounter strange light curves or sources they found difficult to classify, they can post a comment about the source to the project's Talk forum. Once posted, the particularly interesting variables led to considerable discussion. The Zooniverse platform allows any user to search for specific tags, which makes the identification of weird variables relatively easy. There were 330 instances of light curves described as "interesting", 364 described as "unusual" and 92 described as "weird". We show several examples of such variables in Figures 18, 19, and 20. Each of these variables was extensively discussed in the Talk forum because of their bizarre light curves. Table 4 shows the ML classification breakdown for each variable shown in this section.

Many users flagged light curves that have recurrent outliers which indicate the presence of competing sources of variability. These systems are of interest because stars that exhibit multiple pulsation behaviors can act as stellar laboratories, so their identification for additional follow-up is important (Thiemann et al. 2021). An example of such a system is ASASSN-V J085305.34-824360.0 (see Figure 18). On the project, we

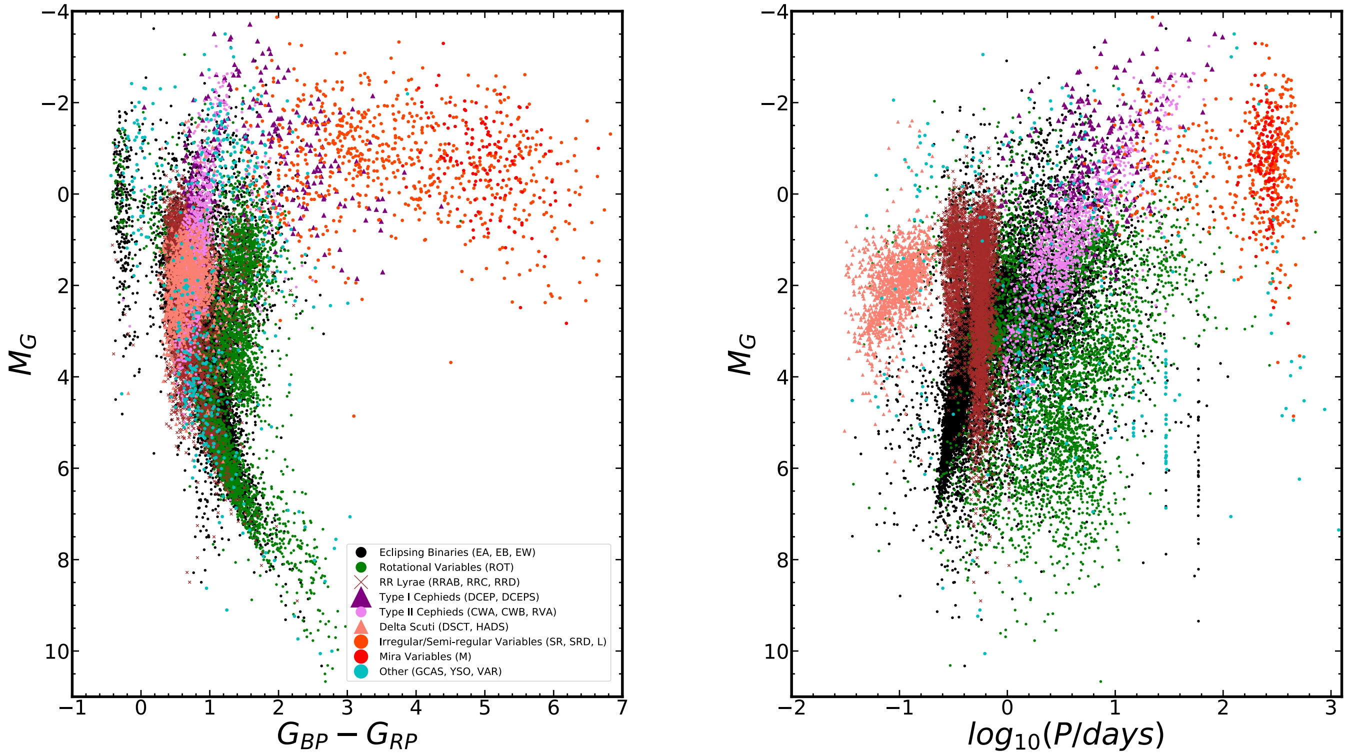


Figure 15. The Gaia EDR3 M_G vs. $G_{\rm BP} - G_{\rm RP}$ color–magnitude diagram (Left) and the M_G vs. $\log_{10}(P/days)$ period–luminosity diagram (Right) for our set of non-Junk variables using labels given by the g-band classifier.

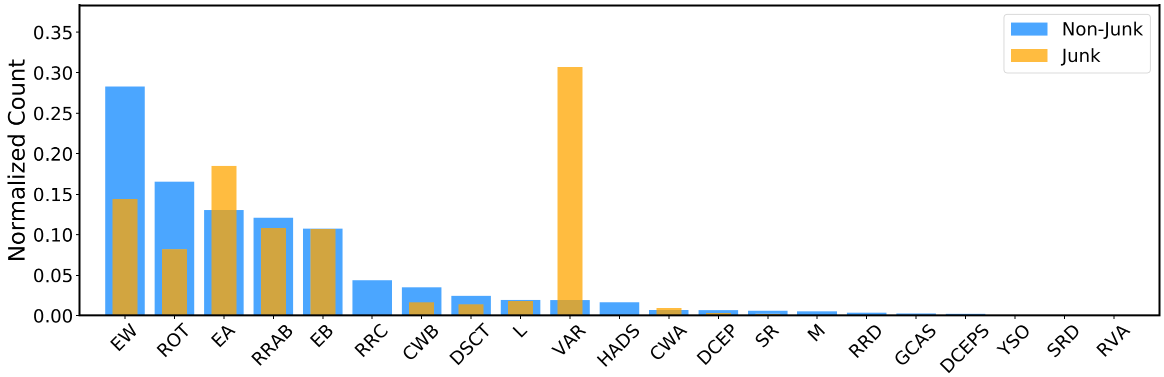


Figure 16. A distribution of the classifications given to the Junk and non-Junk sources by the g-band classifier.

displayed this candidate using the period $P = 10.19$ days. This caught the attention of our users because the light curve displayed a strong sinusoidal variation with recurring outliers at each minima. The regular nature of these outliers indicated that this system might be an eclipsing binary. Additionally, the observed light curve shows distinct amplitude modulation, likely due to spotting on the surface. This star was classified as an eclipsing binary (EB type) in the V-band (ASASSN-V J085305.74-824401.0) with a classification probability of 0.962 and period of 20.4 days. We found the correct period to be $1/2$ of this at $P = 10.21$ days. When phasing the observed light curve with this period, the primary and secondary eclipses become visible, while the rotational signature is blurred. This period is very close to the best GLS period presented to our users on Citizen ASAS-SN, which suggests that this system is a nearly synchronized eclipsing binary with active spotting.

Other particularly interesting and rare systems are pulsating variables in eclipsing binaries. These systems are powerful tools

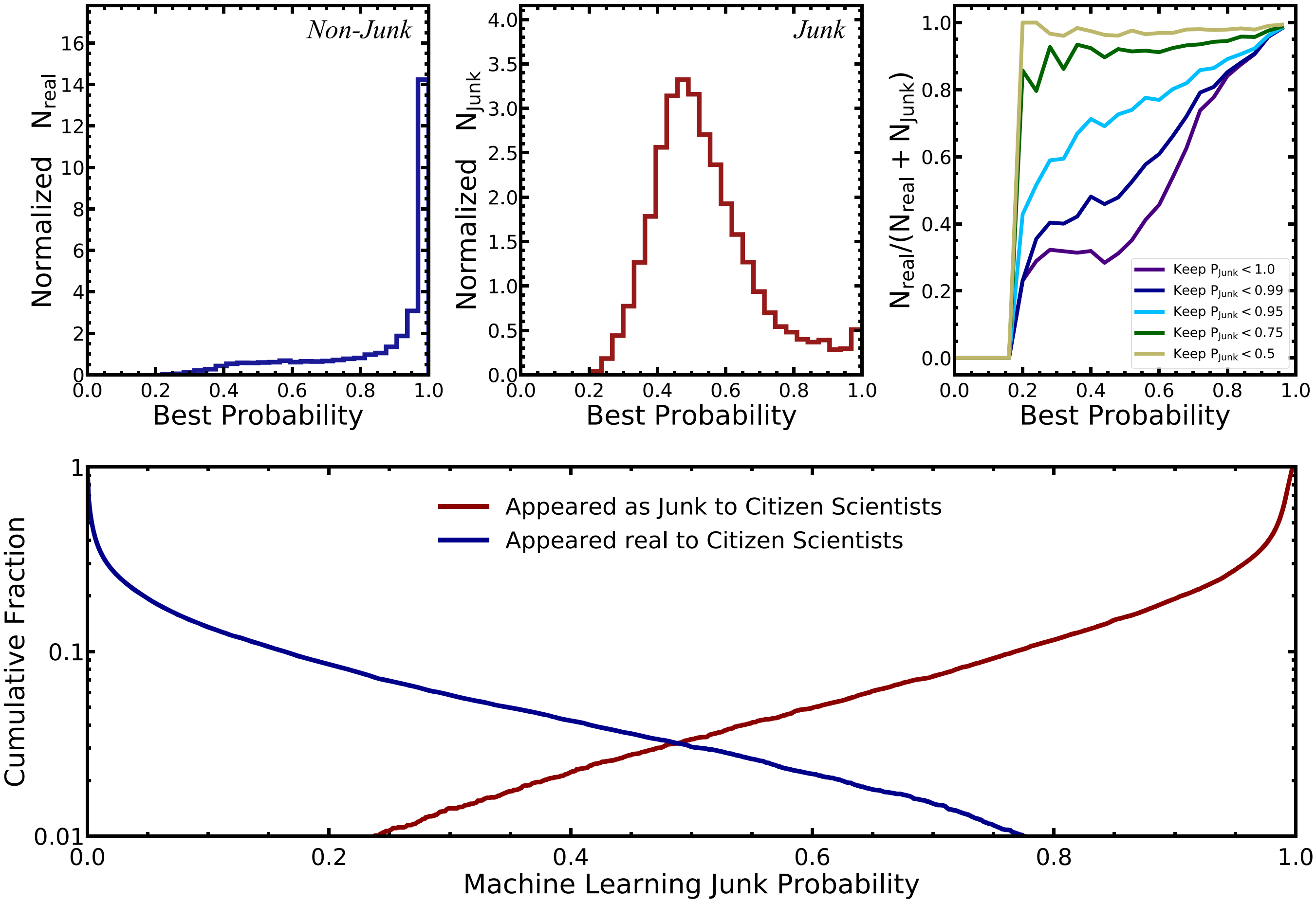


Figure 17. (top, left and middle) Distribution of the best probabilities assigned by the *g*-band classifier to each source in the Junk and non-Junk samples. (top, right) Distribution of the fraction of non-Junk sources viewed as real by the ML classifier as a function of classification probability. (bottom) Distribution of the cumulative fraction of Junk and non-Junk variables as a function of their ML Junk probability.

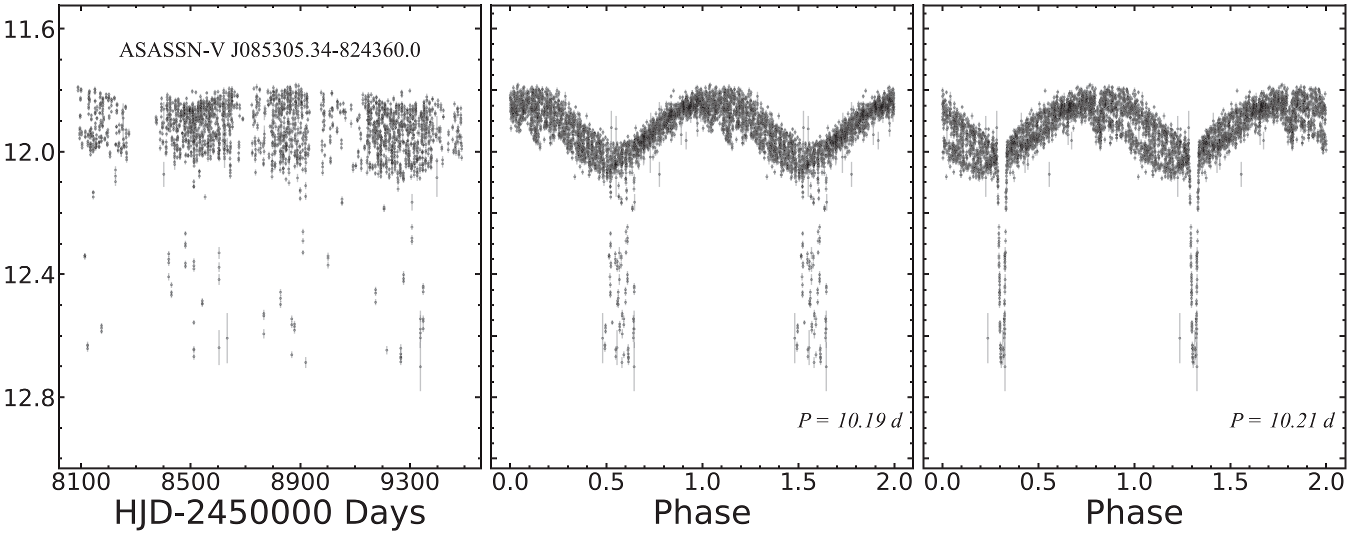


Figure 18. (Left) Observed and phased (Middle and Right) light curves for ASASSN-V J085305.34-824360.0.

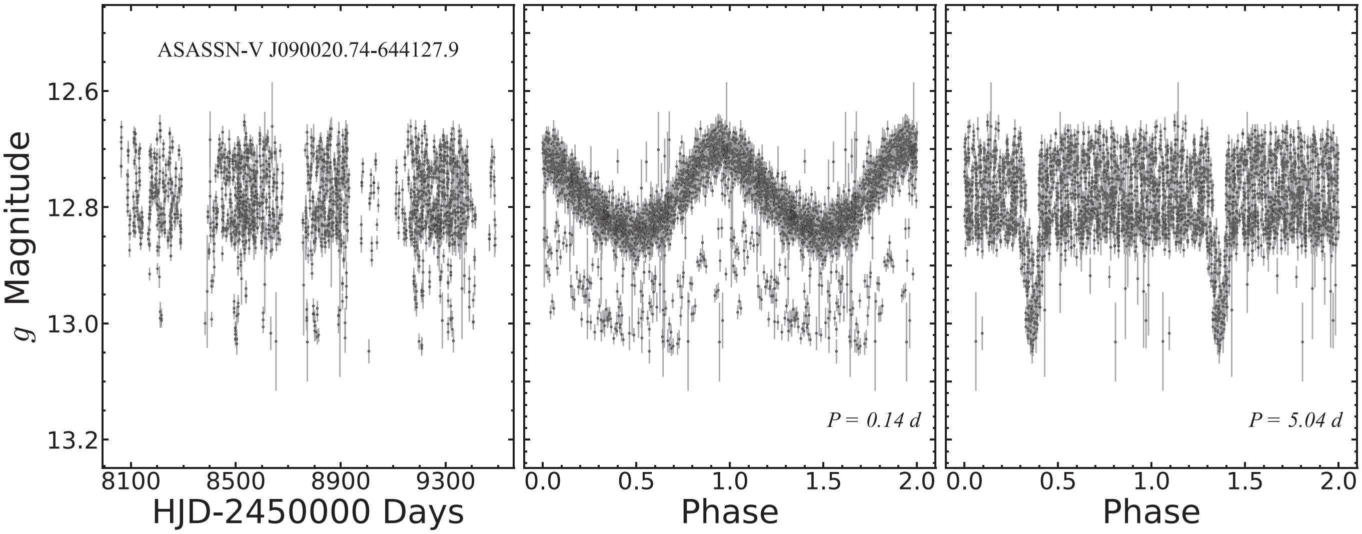


Figure 19. (Left) Observed and phased (Middle and Right) light curves for ASASSN-V J090020.74-644127.9.

Table 3
ML Classification Breakdown of the Real Citizen ASAS-SN Variables

RF Classification	Description	Broad VSX Type	$N_{\rm tot}$	$N_{\rm new}$	$N_{\rm Prob>0.9}$	$N_{\rm Prob>0.5}$
CWA	W Virginis type variables with P > 8 d	Pulsating Variable	206	20	111	182
CWB	W Virginis type variables with P < 8 d	Pulsating Variable	996	15	90	613
DCEP	δ Cephei-type/ classical Cepheid variables	Pulsating Variable	194	35	87	164
DCEPS	First overtone Chepheid variables	Pulsating Variable	64	7	4	46
DSCT	δ Scuti type variables	Pulsating Variable	700	510	348	679
EA	Detached Algol-type binaries	Eclipsing Binary	3729	991	3151	3675
EB	β Lyrae-type binaries	Eclipsing Binary	3074	867	1936	2873
EW	W Ursae Majoris type binaries	Eclipsing Binary	8096	2315	5202	7601
GCAS	Rapidly rotating early type stars	Other	74	27	0	15
HADS	High amplitude δ Scuti type variables	Pulsating Variable	469	195	252	455
L	Irregular Variables	Other	556	155	475	538
M	Mira variables	Pulsating Variable	142	13	141	141
ROT	Spotted Variables with rotational modulation	Rotational Variable	4735	2964	1748	4140
RRAB	Fundamental Mode RR Lyrae variables	Pulsating Variable	3461	1292	2719	3335
RRC	First Overtone RR Lyrae variables	Pulsating Variable	1246	540	897	1185
RRD	Double Mode RR Lyrae variables	Pulsating Variable	103	63	84	100
RVA	RV Tauri variables (Subtype A)	Pulsating Variable	1	1	0	0
SR	Semi-regular variables	Pulsating Variable	172	58	129	164
SRD	Semi-regular variables (Subtype D)	Pulsating Variable	3	0	0	1
YSO	Young Stellar Objects	Other	31	13	4	19
VAR	Variable star of unspecified type	Other	551	339	24	229

because they allow researchers to derive the fundamental stellar parameters and probe the internal structure of stars (Kahraman Aliçavuş et al. 2017). The system ASASSN-V J090020.74-644127.9 is an example where there is a pulsation period of 0.14 days (probably a HADS variable), and a 5.04 day period for the eclipses (see Figure 19). We only presented users with the phased light curves for the pulsation behavior but users pointed out that there may be a hidden eclipse. We believe that the identification of strange variables for additional followup is one of the key strengths of citizen science. A more in-depth analysis of the odd light curves (see Figure 20) and the hybrid systems will be done in future papers.

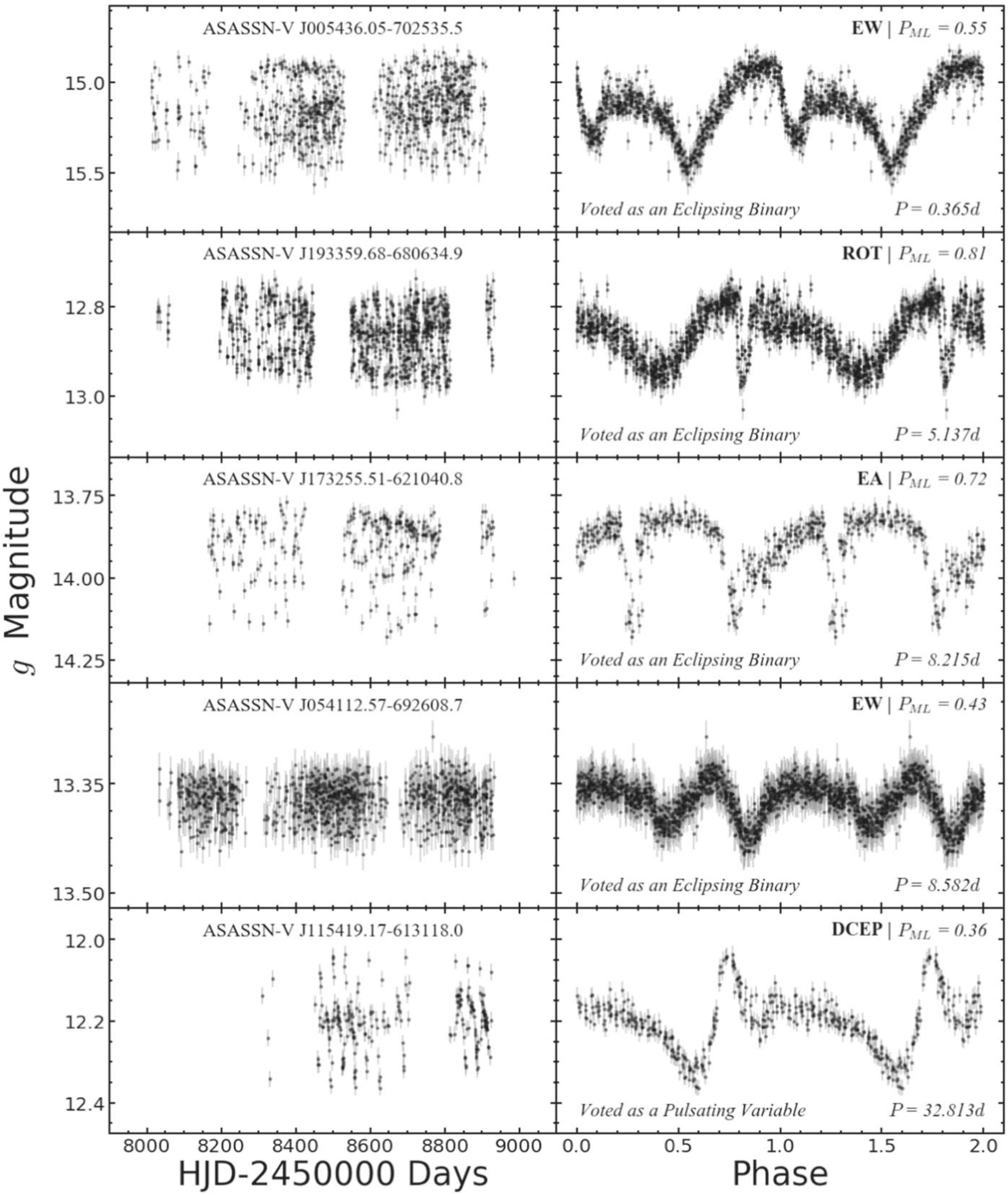


Figure 20. A sample of interesting light curves our users pointed out on the project's Talk forum.

Table 4
Breakdown of the Variable Stars shown in Figures 18, 19, and 20

ID (ASASSN-V)	R.A.	Decl.	$\mathrm{Class}_{\mathrm{CS}}$	P_{CS}	$\mathrm{Class}_{\mathrm{ML}}$	P_{ML}	Period	Other ID
J173255.51-621040.8	263.23129	−62.17799	Eclipsing	0.6	EA	0.725	8.215	ASASSN-V J173255.51-621040.7
J115419.17-613118.0	178.57988	−61.52166	Pulsating	0.5	DCEP	0.364	32.81	ASAS J115419-6131.3
J054112.57-692608.7	85.30236	−69.43574	Eclipsing	0.5	EW	0.431	8.582	OGLE-LMC-ECL-22442
J005436.05-702535.5	13.65020	−70.42652	Eclipsing	0.8	EW	0.553	0.365	WISE J005436.0-702535
J193359.68-680634.9	293.49868	−68.10970	Eclipsing	0.4	ROT	0.811	5.137	ASAS J193359-6806.6
J085305.34-824360.0	133.27224	−82.73333	Eclipsing	0.7	EB	0.658	20.39	ASAS J085305-8244.0
J090020.74-644127.9	135.08642	−64.69109	Pulsating	0.8	EW	0.566	0.289	N/A (New Discovery)

7. Conclusions

We present the first results of Citizen ASAS-SN. This includes the analysis of 403,626 classifications of 40,640 variable candidates at the south celestial pole ($\delta < -60°$) from 2298 users. Classifications for these variables were made between 2021 January 5th and March 27th. The final results are available at the ASAS-SN Variable Star Database (https://asas-sn.osu.edu/variables). The primary classification for each variable comes from the machine learning classifier, but we include the most popular citizen science classification and its probability. Future updates will be released as we move up the sky in decl.. Table 3 lists the number of sources of each variable type in the catalog along with the number of new discoveries in each category.

If we compare the user classifications to either published (VSX, OGLE) or our ML classifications, we found that our volunteers classified eclipsing binaries and pulsating variables most consistently, while struggling to classify rotational variables. We also found that it was exceedingly rare for known variables to be misclassified as Junk, accounting for less than ∼2% for each variable type. We calculated a probability metric for each variable candidate that measures the agreement between users. We found that our users were likely to agree on the classifications for candidates that were most voted as eclipsing binaries, pulsators, and Junk variables. Classifications for variable candidates most voted as rotational and unknown variables were more difficult for our users to agree upon.

User activity generally correlated with higher classification accuracy and higher user agreement, showing that experience improves performance. Our citizen scientists discovered 10,420 new variable sources including, as they defined them, 4234 pulsating variables, 3132 rotational variables and 2923 eclipsing binaries with 131 candidates flagged as Unknown. In addition to these new sources, many users have pointed out unusual or extreme variable candidates on the Citizen ASAS-SN Talk forum for additional follow-up. Moving forward, we plan to release subsequent candidate moving North in decl.. We also plan to extend our workflow to cover higher order classifications including irregular variable stars.

We also built a new g-band machine learning classifier trained on light curves features from variables in our g-band catalog. We found that our classifier was more accurate at classifying known variables than our users. However, the citizen scientists out performed the classifier when it came to identifying Junk light curves. The ML classifier does assign them lower classification probabilities and classifies many of the to the generic VAR class. As built, the ML classifier has to assign all light curves to some type of variable because it has no Junk output class. We now have a Junk training set, and a simple ML classifier simply built to distinguish Junk and non-Junk sources performed very well. We now use this initial Junk classifier to purge these candidates before releasing new light curves to Citizen ASAS-SN. Moving forward, we will rebuild the overall ML variable classifier to include a Junk classification. While it was not one of our initial goals, the construction and continued expansion of a Junk training set will be a very valuable contribution of Citizen ASAS-SN.

We thank the Zooniverse team and each volunteer that participated in Citizen ASAS-SN. We thank the Las Cumbres Observatory and their staff for its continuing support of the ASAS-SN project. We also thank the Ohio State University College of Arts and Sciences Technology Services for helping us set up and maintain the ASAS-SN variable stars database.

The ASAS-SN team at OSU is supported by the Gordon and Betty Moore Foundation through grant GBMF5490 to the Ohio State University, and NSF grant AST-1908570. C.S.K. and K.Z.S. are also supported by NSF grant AST-1814440. Development of ASAS-SN has been supported by NSF grant AST-0908816, the Mt. Cuba Astronomical Foundation, the Center for Cosmology and AstroParticle Physics at the Ohio State University, the Chinese Academy of Sciences South America Center for Astronomy (CAS- SACA), the Villum Foundation, and George Skestos. T.A.T. is supported in part by Scialog Scholar grant 24216 from the Research Corporation. Support for J.L.P. is provided in part by FONDECYT through the grant 1151445 and by the Ministry of Economy, Development, and Tourism's Millennium Science Initiative through grant IC120009, awarded to The Millennium Institute of Astrophysics, MAS. S.D. acknowledges Project 11573003 supported by NSFC. Support for M.P. and O.P. has been provided by the PRIMUS/SCI/17 award from Charles University.

This work has made use of data from the European Space Agency (ESA) mission Gaia (https://www.cosmos.esa.int/gaia), processed by the Gaia Data Processing and Analysis Consortium (DPAC, https://www.cosmos.esa.int/web/gaia/dpac/consortium). Funding for the DPAC has been provided by national institutions, in particular the institutions participating in the Gaia Multilateral Agreement.

This research has made use of the VizieR catalog access tool, CDS, Strasbourg, France. The original description of the VizieR service was published in A&AS 143, 23.

This research made use of Astropy, a community-developed core Python package for Astronomy (Astropy Collaboration, 2013).

Data Availability

The variables are publicly cataloged with the AAVSO and the ASAS-SN light curves can be obtained using the ASAS-SN Sky Patrol (https://asas-sn.osu.edu). The catalog of variables and the associated light curves are available on the ASAS-SN variable stars database (https://asas-sn.osu.edu/variables). The external photometric data underlying this article were accessed from sources in the public domain: Gaia (https://www.cosmos.esa.int/gaia), 2MASS (https://old.ipac.caltech.edu/2mass/overview/access.html), AllWISE (http://wise2.ipac.caltech.edu/docs/release/allwise/) and GALEX (https://archive.stsci.edu/missions-and-data/galex-1/).

ORCID iDs

C. T. Christy https://orcid.org/0000-0003-0528-202X
T. Jayasinghe https://orcid.org/0000-0002-6244-477X
C. S. Kochanek https://orcid.org/0000-0001-6017-2961
T. W.-S. Holoien https://orcid.org/0000-0001-9206-3460

References

Alard, C. 2000, A&AS, 144, 363
Alard, C., & Lupton, R. H. 1998, ApJ, 503, 325
Alcock, C., Allsman, R. A., Alves, D. R., et al. 2000, ApJ, 542, 281
Alhammady, H., & Ramamohanarao, K. 2004, in Fourth IEEE Int. Conf. on Data Mining (ICDM'04), 315
Bellm, E. C. 2014, The Zwicky Transient Facility, arXiv:1410.8185
Brown, A. G. A., Vallenari, A., Prusti, T., et al. 2018, A&A, 616, A1
Brown, T. M., Baliber, N., Bianco, F. B., et al. 2013, PASP, 125, 1031
Clarke, D. 2002, A&A, 386, 763
Derue, F., Marquette, J.-B., Lupone, S., et al. 2002, A&A, 389, 149
Drake, A. J., Djorgovski, S. G., Mahabal, A., et al. 2009, ApJ, 696, 870
Freedman, W. L., Madore, B. F., Hatt, D., et al. 2019, ApJ, 882, 34
Hasanzadeh, A., Safari, H., & Ghasemi, H. 2021, MNRAS, 505, 1476
Heinze, A. N., Tonry, J. L., Denneau, L., et al. 2018, AJ, 156, 241
Holoien, T.-S., Stanek, K. Z., Kochanek, C. S., et al. 2016, MNRAS, 464, 2672
Jayasinghe, T., Kochanek, C. S., Stanek, K. Z., et al. 2018, MNRAS, 477, 3145
Jayasinghe, T., Kochanek, C. S., Stanek, K. Z., et al. 2020, arXiv:2006.10057
Jayasinghe, T., Kochanek, C. S., Stanek, K. Z., et al. 2021, MNRAS, 503, 200
Jayasinghe, T., Stanek, K. Z., Kochanek, C. S., et al. 2019, MNRAS, 485, 961
Jayasinghe, T., Stanek, K. Z., Kochanek, C. S., et al. 2019a, MNRAS, 486, 1907
Jayasinghe, T., Stanek, K. Z., Kochanek, C. S., et al. 2019b, MNRAS, 491, 13
Kahraman Aliçavuş, F., Soydugan, E., Smalley, B., & Kubát, J. 2017, MNRAS, 470, 915
Kochanek, C. S., Shappee, B. J., Stanek, K. Z., et al. 2017, PASP, 129, 104502
Kozlowski, S., Udalski, A., Wyrzykowski, L., et al. 2013, Supernovae and Other Transients in the OGLE-IV Magellanic Bridge Data, arXiv:1301.3909
Lafler, J., & Kinman, T. D. 1965, ApJS, 11, 216
Leavitt, H. S. 1908, AnHar, 60, 87
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2018, Scikit-learn: Machine Learning in Python, arXiv:1201.0490
Pojmanski, G. 2002, The All Sky Automated Survey. Variable Stars in the 0h—6h Quarter of the Southern Hemisphere, arXiv:astro-ph/0210283
Poleski, R., Soszyński, I., Udalski, A., et al. 2012, The Optical Gravitational Lensing Experiment, The Catalog of Stellar Proper Motions toward the Magellanic Clouds, arXiv:1203.2649
Prusti, T., de Bruijne, J. H. J., Brown, A. G. A., et al. 2016, A&A, 595, A1
Riess, A. G., Casertano, S., Yuan, W., et al. 2018, ApJ, 861, 126
Scargle, J. D. 1982, ApJ, 263, 835
Shappee, B. J., Prieto, J. L., Grupe, D., et al. 2014, ApJ, 788, 48
Thiemann, H. B., Norton, A. J., Dickinson, H. J., McMaster, A., & Kolb, U. C. 2021, MNRAS, 502, 1299
Tonry, J. L., Denneau, L., Flewelling, H., et al. 2018a, ApJ, 867, 105
Tonry, J. L., Denneau, L., Heinze, A. N., et al. 2018b, PASP, 130, 064505
Torres, G., Andersen, J., & Giménez, A. 2009, A&ARv, 18, 67
Trouille, L., Lintott, C. J., & Fortson, L. F. 2019, PNAS, 116, 1902
Udalski, A. 2004, The Optical Gravitational Lensing Experiment. Real Time Data Analysis Systems in the OGLE-III Survey, arXiv:astro-ph/0401123
Watson, C. L., Henden, A. A., & Price, A. 2006, SASS, 25, 47
Woźniak, P. R., Vestrand, W. T., Akerlof, C. W., et al. 2004, AJ, 127, 2436
Zechmeister, M., & Kürster, M. 2009, A&A, 496, 577