Publications of the Astronomical Society of the Pacific, 136:084203 (14pp), 2024 August https://doi.org/10.1088/1538-3873/ad68a4

An Automated Catalog of Long Period Variables using Infrared Lightcurves from Palomar Gattini-IR

Aswin Suresh[1], Viraj Karambelkar[2], Mansi M. Kasliwal[2], Michael C. B. Ashley[3], Kishalay De[4,10], Matthew J. Hankins[5], Anna M. Moore[6], Jamie Soon[6], Roberto Soria[7,8,9], Tony Travouillon[6], and Kayton K. Truong[2]
[1] Department of Physics, Indian Institute of Technology Bombay, Mumbai 400076, India; aswin.suresh@iitb.ac.in
[2] Cahill Center for Astrophysics, California Institute of Technology, 1200 E. California Blvd. Pasadena, CA 91125, USA
[3] School of Physics, University of New South Wales, Sydney NSW 2052, Australia
[4] MIT-Kavli Institute for Astrophysics and Space Research, 77 Massachusetts Ave., Cambridge, MA 02139, USA
[5] Arkansas Tech University, Russellville, AR 72801, USA
[6] Research School of Astronomy and Astrophysics, Australian National University, Canberra, ACT 2611, Australia
[7] College of Astronomy and Space Sciences, University of the Chinese Academy of Sciences, Beijing 100049, People's Republic of China
[8] INAF-Osservatorio Astrofisico di Torino, Strada Osservatorio 20, I-10025 Pino Torinese, Italy
[9] Sydney Institute for Astronomy, School of Physics A28, The University of Sydney, Sydney, NSW 2006, Australia
Received 2024 May 9; accepted 2024 July 29; published 2024 August 20

Abstract

Long Period Variables (LPVs) are stars with periods of several hundred days, representing the late, dust-enshrouded phase of stellar evolution in low to intermediate mass stars. In this paper, we present a catalog of 154,755 LPVs using near-IR lightcurves from the Palomar Gattini-IR (PGIR) survey. PGIR has been surveying the entire accessible northern sky ($\delta > -28°$) in the *J*-band at a cadence of 2–3 days since 2018 September, and has produced *J*-band lightcurves for more than 60 million sources. We used a gradient-boosted decision tree classifier trained on a comprehensive feature set extracted from PGIR lightcurves to search for LPVs in this data set. We developed a parallelized and optimized code to extract features at a rate of ∼0.1 s per lightcurve. Our model can successfully distinguish LPVs from other stars with a true positive rate of 95%. Cross-matching with known LPVs, we find 70,369 (∼46%) new LPVs in our catalog.

Unified Astronomy Thesaurus concepts: Long period variable stars (935); Near infrared astronomy (1093); Asymptotic giant branch stars (2100)

1. Introduction

Stars in the final stages of stellar evolution exhibit periodic variations in their brightness on timescales longer than 100 days, and are hence termed as Long Period Variables (LPVs, Samus' et al. 2017). LPVs are generally categorized into three types: Mira variables, semiregular variables, and OGLE[11] small amplitude red giants (Soszyński et al. 2009; Wray et al. 2004). Miras can show oscillation amplitudes larger than ten magnitudes in the optical bands, partly due to absorption of the optical light by nearby molecules. In the infrared, however, the brightness variation is less pronounced. Semiregular variables, which typically display variable amplitudes and periods, and small amplitude red giants are composed primarily of Asymptotic Giant Branch (AGB) stars or Red Giant Branch (RGB) stars and typically show oscillations with amplitudes less than 0.2 mag (Smak 1966).

The advent of time domain surveys has resulted in a number of variable star catalogs covering a wide range of optical variability in stellar populations. The Massive Compact Halo Object survey (Alcock et al. 2001) found 20,000 LPVs in the Large Magellanic Cloud (LMC). Wood et al. (1999) used data from 1500 LPVs in the LMC to establish several linear sequences in the log(period)–log(luminosity) space of LPVs. OGLE (Udalski et al. 1992) identified 91,995 LPVs in the LMC from the OGLE-III catalog of variable stars by examining the log(period)–log(Wesenheit index) relation, finding 1667 Miras in the LMC (Soszyński et al. 2009). The All Sky Automated Survey for Supernovae (ASAS-SN, Shappee et al. 2014) optical survey monitors the entire sky with a cadence of less than a day to a depth of 18.5 g mag. Jayasinghe et al. (2019) classifies 412,000 known variable stars using ASAS-SN *V*-band lightcurves by training a random forest classifier on period, color, and Fourier fit-based features with high accuracy. They provide ∼300,000 definite classifications, of which 92,230 are semi-regular variables, and 10,239 are Mira variables. *G*-band lightcurves from the survey were also used

[10] NASA Einstein Fellow.
[11] Optical Gravitational Lensing Experiment (Udalski et al. 1992).

to discover 116,000 new variable stars, again using a random forest classifier trained primarily on color and magnitude-based features. They find 363 new Mira variables and 57,925 new semi-regular variables (Christy et al. 2023). The All Sky Automated Survey (Pojmanski 2002) carried out a two-year *I*-band survey covering 300 sq. deg. of the sky. It found 3900 variable stars by searching for stars showing high standard deviations in their lightcurves and manually removing spurious cases. Over 3500 objects in the catalog under the tag "miscellaneous" include Mira variables and irregular and multi-periodic pulsators (Pojmanski 2002).

The Vista Variables in the Vía Láctea (VVV) survey (Dekany et al. 2011) discovered 129 Miras in the Galactic bulge, along with producing a catalog of 1013 LPVs in the Milky Way using point-spread function photometry data in the K_s band from 10 tiles around the Galactic bulge and studied the period–amplitude relationship of Miras in the near-infrared (NIR) (Nikzat et al. 2022). The largest catalog of LPVs to date, consisting of over 1.7 million LPV candidates, of which 392,240 have published periods, was presented in Lebzelter et al. (2023) using 34 months of data from the third Gaia data release. Karambelkar et al. (2019) identified 417 extremely luminous LPVs in nearby galaxies from the SPitzer InfraRed Intensive Transients Survey (Kasliwal et al. 2018) with the Spitzer Space Telescope and extended the period–luminosity relationship of AGB stars to longer periods and luminosities. Several catalogs of variable stars consisting of a mixture of short-period ($<$100 days) stars, such as Cepheids and RR Lyrae, as well as longer-period variables, have been identified (see for e.g., Pojmanski et al. 2005; Chen et al. 2018, 2020). The wide parameter space of stellar variability and increasingly large data sets from time-domain surveys present the biggest challenge in these studies to identify and characterize variable stars.

The efficacy of Machine Learning (ML) techniques in identifying and classifying variable stars has been demonstrated in multiple works. For instance, Masci et al. (2014) used a random forest classifier trained on lightcurve morphology-based features and Fourier decomposition to distinguish between Algols, RR Lyrae, and W Ursae Majoris among 8273 variables from WISE. Debosscher et al. (2007) employ a multivariate Gaussian mixture classifier to classify between 35 classes of variable stars solely using Fourier amplitudes and phases as the feature set. Sánchez-Sáez et al. (2021) classified 868,371 sources using lightcurves from the Zwicky Transient Facility (Bellm et al. 2019) using a multi-level classifier architecture trained on 152 features. Boone (2019) developed a gradient-boosted decision tree classifier trained on features extracted from Gaussian process augmented multi-band lightcurves to classify between several variable and transient classes ranging from Miras and RR Lyrae to TDEs, SNe Ia, and kilonovae.

In this paper, we employ an ML approach to search for LPVs using data from the Palomar Gattini-IR (PGIR, De et al. 2020) survey taken between 2018 September and 2021 July. PGIR is a wide-field, NIR robotic time domain survey operating at Palomar Observatory, California. Operating in the NIR *J*-band, PGIR has a field of view of 24.7 sq. deg (4.96 deg $\times$ 4.96 deg) with a pixel scale of $8\farcs7\,\mathrm{pixel}^{-1}$. The images are then resampled to provide a pixel scale of $4\farcs3\,\mathrm{pixel}^{-1}$ and a PSF FWHM of 1.2–2.1 pixels on the science frames. Further details about the survey strategy and data processing can be found in De et al. (2020). PGIR's survey of the northern sky (decl. $> -28°$) at a cadence of $\approx$2–3 days to a depth of $\sim$16 AB mag, has produced J-band lightcurves of more than 60 million stars. We utilize a ML framework to identify LPVs in this lightcurve archive. We build a comprehensive training set consisting of LPVs, other erratic and non-periodic variables such as RCrB stars, Young Stellar Objects and non-variables. We train a gradient-boosted decision tree to distinguish between these classes, and use its predictions to generate a catalog of 159,696 LPVs. Of these, 73,346 (45.9% of the full catalog) are newly identified LPVs. We describe the training sample and feature set for the ML classifier in Section 2. We explain the training procedure, iterative build-up of the training set, and the relative importance of various features in Section 3. We analyze the features of LPVs in the PGIR catalog and assess the validity, completeness, and purity of the catalog by comparison with other catalogs in Section 4. Section 5 summarizes our conclusions.

2. Training Sample

The default data processing pipeline of PGIR (De et al. 2020) produces light curves for sources detected in multi-epoch images as part of "match files." Briefly, the reference images of each field is used to construct a master source catalog, and spatially cross-matched to the source catalog for every observation of the field as part of survey operations. The resulting cross-match produces light curves for every source detected in the reference image covering the epochs where the source is detected in the individual epochs. In this paper, we use data from the beginning of PGIR survey operations until 2021 July 14, giving a baseline of $\sim$1400 days for each lightcurve, which is well-suited for characterizing the few hundred day periods of LPVs.

We note that spatially cross-matching sources from the reference image to those in science images may occasionally result in false associations due to varying image depths and seeing. These cases will show up as unphysical fluctuations in the lightcurve. To minimize the effect of these on our catalog, we resample the lightcurves using Gaussian Process fits (see Section 2.2), which discards a majority of such outliers. Additionally, the PGIR reference and science images have

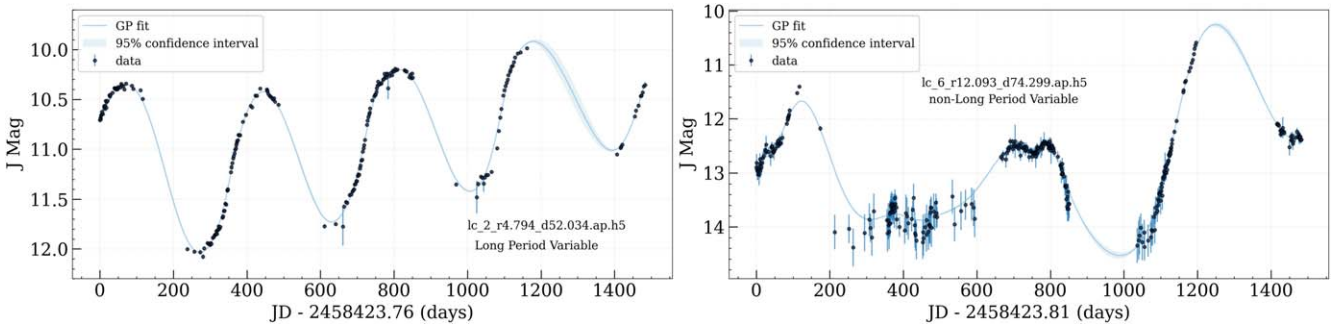

Figure 1. Examples of LPVs and non-LPVs from the bonafide set. LPVs display an approximately sinusoidal morphology in their lightcurves, as can be seen in the figure on the left, while non-LPVs show erratic behavior, as the spectroscopically confirmed RCorBor variable shown on the right. The blue line and shaded blue region represent the mean of the Gaussian process regression fit and the 95% confidence interval respectively.

good astrometric similarity as PGIR observes on a fixed sky grid, which further minimizes false association errors.

Constructing a catalog of LPVs requires distinguishing LPVs from non-periodic variable stars (referred to as "non-LPVs" hereafter) and non-variable stars ("non-vars" hereafter) with high accuracy. The vast parameter space spanned by these stellar lightcurves makes ML a viable option to learn intricate features distinguishing the three classes. The starting point for building the ML classifier is assembling a training data set representative of the various characteristics of LPVs, non-LPVs and non-vars. We build this training set iteratively. We first train a classifier on a small set of bonafide members of each category. We use this classifier to predict the classes of a larger sample of sources, visually examine the results and use the correctly classified sources to expand the initial training set. The initial training set is described in Section 2.1 and the final, expanded data set is described in Section 3.2.

2.1. Initial Bonafide Training Set

We begin with a relatively small training set of 1344 sources, consisting only of LPVs and non-LPVs. These were assembled as part of ongoing searches for the erraticaly varying R Coronae Borealis (R Cor Bor, Clayton 2012) stars and other erratic, large-amplitude Galactic variables using PGIR (Karambelkar et al. 2021; N. Earley et al. 2024, in preparation). Our "LPV" set includes stars identified as LPVs in these studies, primarily based on visual inspection of their lightcurves and Lomb Scargle (Lomb 1976; Scargle 1982) periodograms. We also include any additional late M-type stars that were spectroscopically classified as part of these searches. The "non-LPV" set consists of all known, spectroscopically confirmed R Cor Bor stars that show erratic variations in PGIR data and a smattering of other erratically varying stars such as young stellar objects. LPVs constitute the majority sample of this initial training set, with 1265 samples, and non-LPVs form the minority class with 79 samples (5.87% of the full data set). Crossmatching the initial training set with SIMBAD, we obtain 1313 objects with classifications. Of these, 1296 fall into the desired categories for LPVs/non-LPVs. Representative examples of the two classes are shown in Figure 1.

This training sample may not be sufficient to cover the breadth of parameters spanned by the two classes. Hence it is used as a baseline data set to obtain broad classifications. The classifier will inevitably have a high misclassification rate in the first few training iterations before a more balanced data set enables it to learn the characteristics of both classes. To mitigate this, we built a bigger data set iteratively by visually examining the predictions of previous classifiers and re-labeling misclassified sources manually (see Section 3.2).

2.2. Feature Extraction

We use a set of 16 features for initial training iterations. We begin by fitting a Gaussian Process (GP; Williams & Rasmussen 1995) to model the lightcurves. Gaussian processes have been shown to be an effective method for astronomical lightcurve modeling (see for e.g., Foreman-Mackey et al. 2017; Boone 2019; Qu et al. 2021). We use the python package `george` (Ambikasaran et al. 2015) to implement the GP fits. The kernel of choice is a Matern 3/2 kernel with a length scale set to the period extracted from the raw lightcurve. A uniform time sampling of 500 points from the first observation to the most recent observation is used to calculate the predicted *J*-band magnitudes. The minimum and maximum slopes and peak-to-peak amplitudes are calculated from the augmented lightcurve. The goodness of fit of the Gaussian process is quantified by the R^2 score and is saved as the "GP score." We calculate the difference between the 90th percentile and 50th percentile of the *J*-band magnitudes of each lightcurve as the percentile difference.

Periodicity-based features are extracted using the Lomb–Scargle periodogram, implemented using the `gatspy` library (VanderPlas & Ivezić 2015). The augmented lightcurve, with *J* band magnitude, converted to flux, is fit to obtain the three

most prominent periods and their associated Lomb–Scargle score. A sinusoid with the best fit period is fit to the lightcurve to calculate the reduced χ^2 of the fit. Additionally, the reduced null χ^2 is calculated as

$$\chi^2_{\mathrm{red,null}} = \sum_i \left(\frac{y_i - \hat{y}}{\sigma_i} \right)^2 \quad (1)$$

where y_i and σ_i are the fluxes and errors respectively, and $\hat{y}$ is the error weighted mean of J band flux.

The final periodicity-based features used from these calculations are the Lomb–Scargle score of the best period—LS score, reduced χ^2 and reduced null χ^2, ratio of the best fit period to the maximum period (defined as the observation baseline)—period ratio, ratio of Lomb–Scargle score of the two most prominent periods—Lomb–Scargle score ratio and ratio of reduced χ^2 and reduced null χ^2–χ^2 ratio. Phase-folded lightcurves of LPVs show a clean sinusoidal evolution. The χ^2 value of a sinusoidal fit to the phase folded lightcurve is saved as "phase χ^2" and is useful to distinguish non-LPVs, non-variables, and erratic variables, which have high phase χ^2 values compared to LPVs.

Another useful measure of variability is given by the Stetson L index (Stetson 1996):

$$L = \frac{JK}{0.789}, \quad (2)$$

where

$$J = \sum_{n=1}^{N-1} \mathrm{sign}(\delta_n \delta_{n+1}) \sqrt{|\delta_n \delta_{n+1}|}, \quad (3)$$

and

$$K = \frac{1/N \sum_{i=1}^{N} |\delta_i|}{\sqrt{1/N \sum_{i=1}^{N} \delta_i^2}}, \quad (4)$$

are the Stetson J and K indices (Stetson 1996), with δ_i being the difference between the magnitude of the ith point and the mean magnitude of the lightcurve, scaled by the error of the ith point.

The final feature extracted from lightcurves is the von Neumann score (von Neumann 1941) calculated as

$$\eta = \frac{\sum_{n=1}^{N-1} (x_{n+1} - x_n)^2 / (N-1)}{\sigma^2}. \quad (5)$$

η is a measure of the correlation between successive observations, and positive correlations lead to small values of η.

Using a fully parallelized and optimized feature extraction process, our code can calculate all features for each lightcurve in ∼0.1 s. The distribution of select features for LPVs and non-LPVs in the bonafide set is shown in Figure 2.

We supplement the features obtained directly from PGIR lightcurves using color information of variable stars observed by other ground and space-based missions. In particular, we query the J, H, and K band colors of variables from the

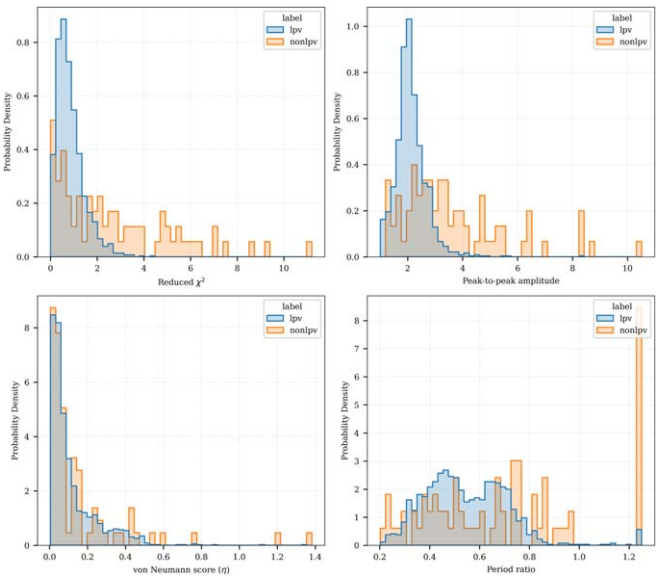

Figure 2. Distributions of select features for variables in the bonafide set, shown as probability densities. Features such as the Reduced χ^2 and the peak-to-peak amplitude can be seen to separate the LPVs and non-LPVs while the period ratio and the von Neumann score occupy similar values for both classes.

2-Micron All Sky Survey (2MASS) (Skrutskie et al. 2006). The 2MASS All-Sky Catalog of Point Sources (Cutri et al. 2003) contains over 300 million objects, including variable stars. We query the infrared photometry of crossmatched sources and calculate the J–H and J–K colors. We note that the J–H and J–K colors are highly correlated in the bonafide set. The Wide-field Infrared Survey Explorer (WISE; Wright et al. 2010) is a space-based infrared observatory surveying the infrared sky at 3.4 (W1), 4.6 (W2), 12 (W3) and 22 (W4) μm wavelengths. We query the WISE All-sky data release (Cutri et al. 2012) to obtain W1–W2, W1–W4, and W3–W4 colors. For cases where no crossmatch is present in WISE or 2MASS, the median color value of the entire data set is used.

2.3. Data Resampling

To overcome the class imbalance in the initial training set, we use synthetic sampling methods implemented using the `imbalanced-learn` package (Lemaître et al. 2017). Multiple strategies exist to selectively upsample the minority class and undersample the majority class. Based on trial and error, we choose the Adaptive Synthetic (ADASYN) sampling method (He et al. 2008) for upsampling and all K-Nearest Neighbours (allKNN) method (Tomek 1976) for undersampling.

ADASYN sampling creates new synthetic samples for the minority class by interpolating between existing minority samples. This interpolation is performed by considering the neighboring minority samples and creating synthetic samples that lie along the line connecting them. ADASYN is unique in

the sense that it adjusts the number and distribution of synthetic samples based on the difficulty of classification. It focuses on generating synthetic instances in the regions of the feature space that are challenging for the classifier. By doing so, it helps to balance the class distribution while maintaining the overall integrity of the data, resulting in improved model performance on imbalanced data sets.

The allKNN downsampling algorithm works by assigning a minority density to each sample in the minority class. The minority density is calculated as the number of minority samples in k-nearest neighbors, where "k" is a user-controlled parameter. Samples from the majority class are selectively removed if they fall into a minority neighborhood with a density below a set threshold. Samples with density below the threshold could be sparsely located or overlap with the majority class, making it hard for the model to learn. This process is iterated untill a desirable balance is achieved. allKNN downsampling, therefore helps create a balanced data set while also removing samples that are harder to learn.

This combination is used to obtain a more balanced training set used to train the classifier.

3. Machine Learning Framework

Decision trees and random forest-based classifiers have been shown to be effective for the task of classifying variable star lightcurves (see for e.g., Debosscher et al. 2007; Masci et al. 2014; Boone 2019; Sánchez-Sáez et al. 2021). Decision trees (Breiman et al. 1984) are intuitive models having a tree-like branched node structure. They recursively partition the input data at each node to create homogeneous subgroups based on criteria set from the features. Data can be split either to maximize information gain or minimize impurity in partitioned sets. Complex decision trees may be prone to overfitting. Parameters such as the maximum depth of the tree and the minimum number of samples required to create a split can be controlled to avoid overfitting data. Random forests (Breiman 2001) is an ensemble method that builds multiple decision trees and combines their predictions to assign a probability to the predicted output. Different trees in the forest are built from random subsets of the training set and features. This helps avoid overfitting and gives improved performance. The importance of features in separating various classes in the input data can also be quantified in random forests.

We employ a gradient-boosted decision tree classifier implemented using the `LightGBM` framework (Ke et al. 2017). Gradient boosting is a popular technique for creating ensemble decision trees. It sequentially builds weak-learner trees where each tree is trained to correct the residuals of the preceding tree, "boosting" its performance. Gradient descent is used to minimize a loss function, which is how the model learns. `LightGBM` offers an efficient implementation of gradient-boosted decision trees and reduces memory consumption by binning features.

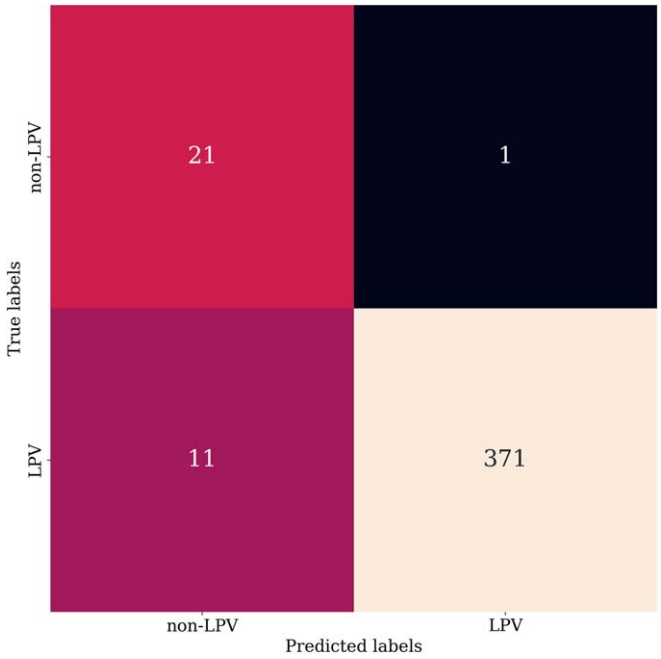

Figure 3. Confusion matrix from training on the bonafide data set. This model will be referred to as LGBMv0.1. Metrics: TPR = 97.120%, TNR = 95.455%, prec = 99.731%, F-measure = 0.984, weighted g-mean = 0.970. The number in each cell of the matrix represents the number of objects classified as the predicted label (x-axis) as opposed to their true label (y-axis).

For the training process, we split the bonafide data set in a 70:30 train-test split. It is important that the model is not tested on resampled data, hence resampling is done only on 70% of the bonafide set. LPVs are labeled as 1 and non-LPVs as 0. Due to class imbalance, the overall accuracy of the classifier may be misleading. We use the following metrics to quantify model performance:

$$\text{Recall / True positive rate (TPR)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

$$\text{True negative rate (TNR)} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{7}$$

$$\text{Precision (prec)} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

$$F\text{ - measure} = \frac{2 * \text{prec} * \text{recall}}{\text{prec} + \text{recall}} \tag{9}$$

$$\text{Weighted } g\text{-mean} = (\text{TPR}^{\text{P}} * \text{TNR}^{\text{N}})^{\frac{1}{\text{P}+\text{N}}}, \tag{10}$$

where TP is true positive, FP is false positive, TN is true negative, FN is false negative, P is all positives, and N is all negatives.

The true positive rate is the probability of an LPV being labeled correctly, and the true negative rate is the probability of

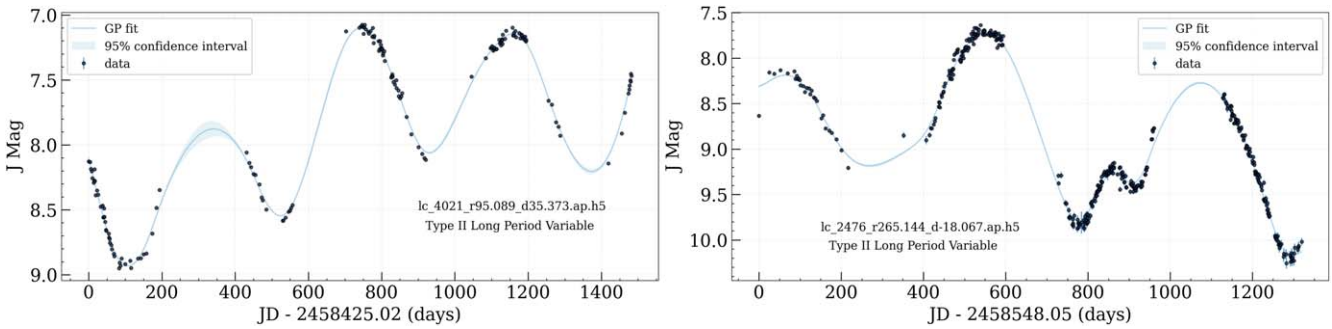

Figure 4. Examples of LPVs classified as Type II LPVs. Type II LPVs display sinusoidal evolution, however, additional trends such as a constant rise or deacy, varying amplitudes of oscillation or presence of several oscillation frequencies are also present.

a non-LPV being labeled correctly. Precision gives the fraction of all LPVs classified correctly. The *F*-measure is an accuracy metric given by the harmonic mean of precision and recall. The *g*-mean is the geometric mean of TPR and TNR. We observe that due to class imbalance in the test set, the simple geometric-mean tends to favor models that have a lower misclassification rate of non-LPVs. To mitigate this bias, we calculate the weighted geometric mean of TPR and TNR, using the total samples in each class as weights to get a robust measure of classifier performance.

3.1. Training on the Bonafide Set

The `LightGBM` gradient-boosted decision tree has a large number of hyperparameters that need to be tuned to get the most out of the classifier. A grid search that chooses the optimal hyperparameters by trying each possible combination would be time-consuming and is not feasible. In each training iteration, we randomly sample 200 points from the hyperparameter space and choose the set with the highest accuracy. This method has been shown to be effective at achieving near-optimal performance with high probability (Bergstra & Bengio 2012). For each combination of hyperparameters, accuracy is calculated using five-fold cross-validation. The training data set is partitioned into five folds, with the model being fit to four such folds and evaluated on the remaining fold. This process is repeated to effectively sample the training data. We evaluate the metrics on the model and create a confusion matrix. For binary classification tasks, the confusion matrix is a 2×2 matrix showing the distribution of true negatives, false positives, false negatives, and true positives. Examining the confusion matrix gives good insights into the model's performance, as all metrics are derived from its quantities. Figure 3 shows the confusion matrix from training on the bonafide data set.

Following this training procedure, it is important to understand samples misclassified by the model. We save the predicted probabilities and labels for all samples in the training set and compare labels with the true values. Lightcurves of all misclassified samples are plotted and saved along with the confusion matrix. Since we perform a randomized search for hyperparameters, we run multiple training iterations and pick the best-performing model. Examination of confusion matrix and misclassified samples is particularly helpful in choosing the optimal model in cases where certain samples may be mislabelled in the training set. This is particularly unavoidable when we build the training set in an iterative process and predictions from initial models form the true labels of the next iteration.

3.2. Expanding the Training Set

The first step in expanding the training data set is to choose a larger set of data to obtain predictions on. We extracted features for 35 million lightcurves in the PGIR lightcurve database. Of these, we choose a subsample of $\sim$12,000 sources by applying cuts on the von Neumann score ($\eta < 1.5$), peak-to-peak amplitude (ptp > 1) and Lomb–Scargle score (LSscore > 0.75). We run the LGBMv0.1 classifier on all of these sources and choose only those samples with a prediction probability >0.9, resulting in a larger training set of 9815 LPVs and 443 non-LPVs.

However, upon visual inspection of the classifications, we find that the misclassification rate is high in this data set. We observe that a large fraction of the classified non-LPVs are LPVs that show variations in amplitude, show multiple pulsation modes or have a slow background decay or rise in addition to the few hundred day periodic variations. Upon training on the extended data set with corrected labels, we observe that the classifier struggles to separate these erratic LPVs, generally deviating from a clean sinusoidal evolution, from non-LPVs. Hence we create a third category—"Type II LPVs," based purely on lightcurve morphology comprising of such non-sinusoidal LPVs. We note that this classification does not necessarily hold a physical meaning, as many AGB stars typically do display such evolution and can have multiple simultaneous periodic excitations. However, separating the LPVs into cleanly sinusoidal and erratic evolution better

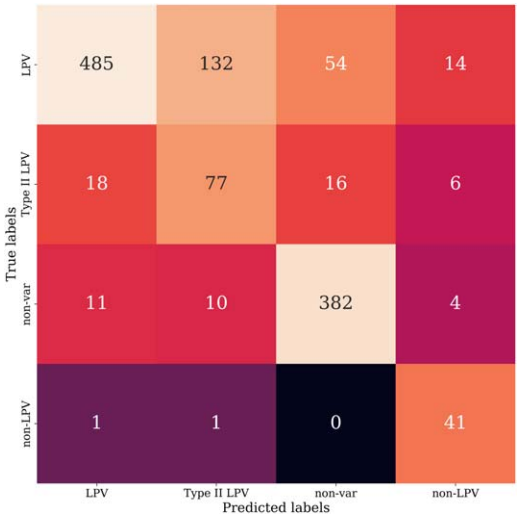

Figure 5. Confusion matrix from training on the extended labeled set. This model will be referred to as LGBMv0.2. Since we classify between three classes, we calculate metrics for each class. Metrics (listed for LPVs, Type II LPVs, non-vars and non-LPVs respectively): TPR = 70.80%, 65.81%, 93.86%, 95.35%, TNR = 94.71%, 87.40%, 91.72%, 98.02%, precision = 94.12% 35.0%, 84.51%, 63.08%, F-measure = 0.81, 0.46, 0.89, 0.76 and weighted g-mean = 0.80, 0.69, 0.93, 0.95.

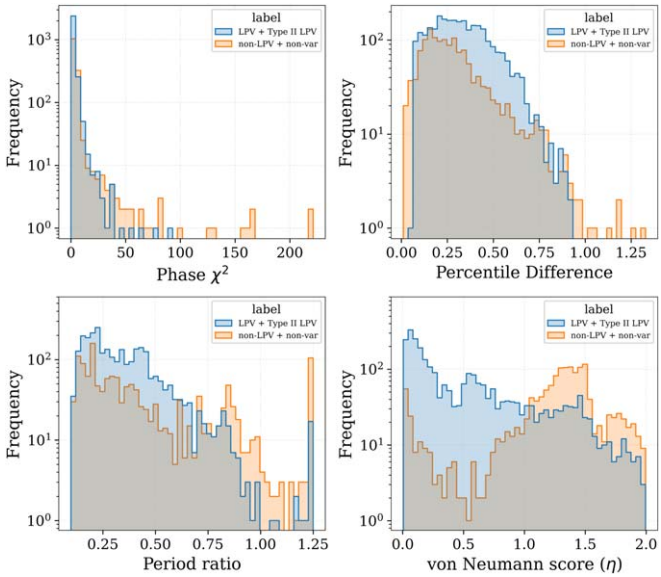

Figure 6. Distributions of select features for variables in the expanded training set. We observe that the features occupy similar values for the classes shown here as the classifier has to separate low amplitude variables whose feature phase space is not too different from non-variables.

enables the classifier to pick up the general features of each class. We combine the sinusoidal LPVs and type II LPVs into the LPV class following classification. Representative examples of type II LPVs are shown in Figure 4.

We repeat the training procedure on this larger sample of labeled data. In this iteration, we observe that the classifier has improved performance and is better able to distinguish the type II LPVs and non-LPVs, as shown in the confusion matrix in Figure 5. We assemble a clean set of non-vars by visually inspecting a subset of lightcurves chosen to have von Neumann score less than 1.5, amplitude less than 1 mag and LPV probability less than 0.3 from the classifier trained on bonafide variables. Finally, we build a master training set consisting of 2335 LPVs, 444 Type-II LPVs, 1332 non-vars and 166 non-LPVs. The distribution of select features for LPVs and non-LPVs in the bonafide set is shown in Figure 6. Several training iterations of the model on this training set resulted in a best-performing model with a g-mean score of 0.95.

The model shows high confusion between LPVs and Type II LPVs, but is able to minimize confusion between the combined set of these two and the combined set of non-vars and non-LPVs. We sum the probability score of LPV and type II LPV is to calculate the net probability of a source being an LPV. This allows us to separate LPVs from non-LPVs and non-vars in an efficient manner. We group all of the type II LPVs into the LPV class and the non-vars into the non-LPV class to obtain the final confusion matrix and metrics shown in Figure 7. We also list the optimal parameters of the gradient-boosted classifier in Table 1.

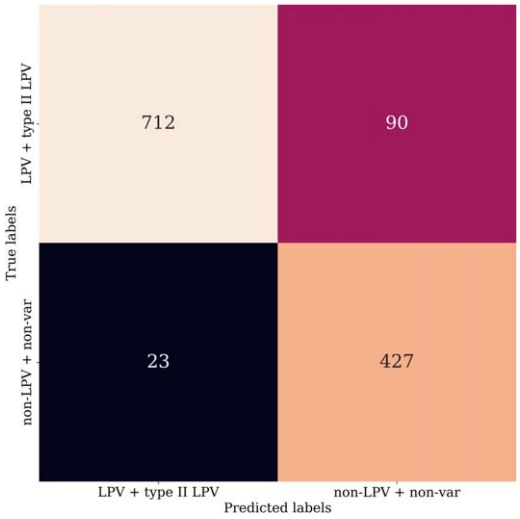

Figure 7. Confusion matrix from training on the extended labeled set with Type II LPVs grouped with LPVs and non-LPVs grouped with non-vars. Metrics: TPR = 94.89%, TNR = 88.78%, precision = 82.56%, F-measure = 0.88, weighted g-mean = 0.95.

3.3. Feature Importance

Tree-based classifiers provide convenient methods to quantify the relative importance of different features used in training. The `LightGBM` framework, in particular, calculates feature importance in two ways, using "split" and "gain." Split-wise importance is calculated as the number of times a particular feature is used in the

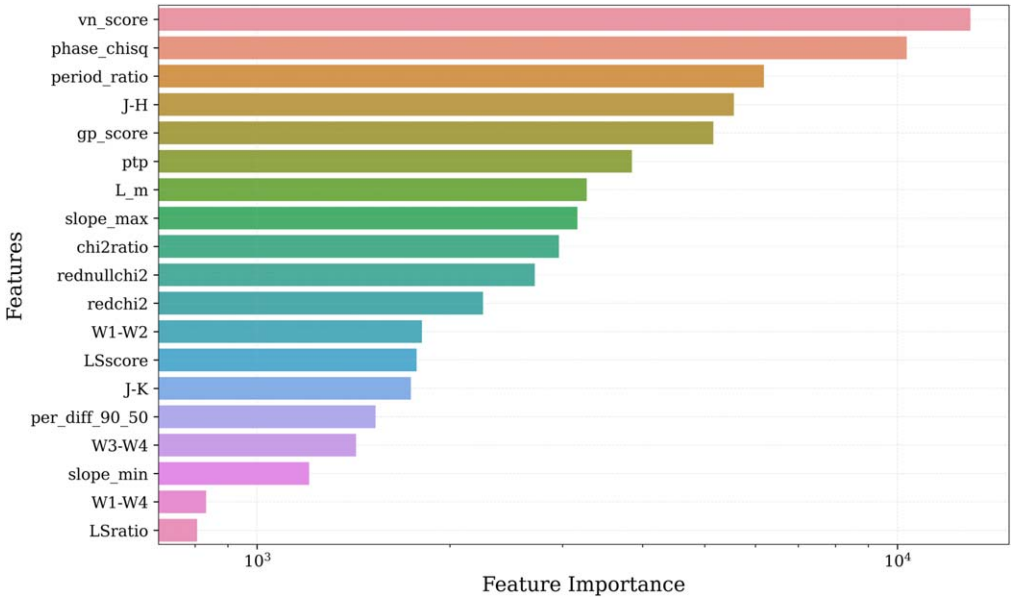

Figure 8. Feature importance of various features used in training. The feature importance quantifies how effective a particular feature is in separating an LPV and a non-LPV. We notice features such as the von Neumann score and period ratio are very effective, while the Lomb–Scargle score is not as useful for the classifier.

Table 1
Optimal Hyperparameters of the `LightGBM` Classifier

Hyperparameter	Value
`objective`	`multiclass`
`boosting_type`	`gbdt`
`colsample_bytree`	0.9
`max_depth`	9
`min_child_samples`	40
`min_split_gain`	1
`n_estimators`	460
`num_leaves`	197
`reg_alpha`	0.035
`reg_lambda`	0.03
`subsample`	0.90
`subsample_freq`	59

model to partition data. Since the model tries to split data to minimize a loss function, the number of times a feature is used in this process is a good measure of its effectiveness. Gain-wise importance calculates the impurity reduction in each split; impurity reduction is the difference in loss of parent tree node and the weighted sum of loss of child tree nodes after the split. This quantifies how much a split improves the model's performance. Impurity reduction for each feature is accumulated for all trees in the ensemble and normalized to obtain the final relative importance. We adopt gain-based importance as the preferred metric as it provides a direct measure of improvement in accuracy contributed by a feature.

Figure 8 shows the calculated feature importance. We note that the period ratio is among the most important features for the classifier through both the gain and split importance calculations. This is consistent with other variable star or transient alert classifiers (e.g., Kim & Bailer-Jones 2016; Masci et al. 2017; Sánchez-Sáez et al. 2021). The χ^2 of a sinusoid fit to the phase folded lightcurve and the *J–H* color also rank highly in both metrics, while the Stetson *L* index and peak-to-peak amplitude perform decently. We also observe that the Lomb–Scargle scores and percentile difference perform poorly in separating LPVs and non-LPVs.

4. The Catalog

We use the LGBMv0.2 classifier to predict the classes of 35 million PGIR lightcurves, chosen to have a von Neumann score less than 2, and requiring at least 30 detections. We present a catalog of 154,755 stars having *prob_lpv_sum* > 0.5 from the PGIR survey in Table 2, containing the following columns. "Name" refers to the internal designation of the variable, "R.A." and "decl." give the spatial position, "num_detections" is the number of detections, "lcdur" is the baseline of observation, given by the difference of the time of final detection and the time of initial detection, "bestperiod" and "amplitude" are the period and amplitude, "J_mag" is the mean *J* band magnitude, "prob_lpv_sum" is the classifier confidence. In addition, all extracted features and 2MASS and WISE color information are also presented. Figure 9 shows example lightcurves of 20 LPVs from the catalog.

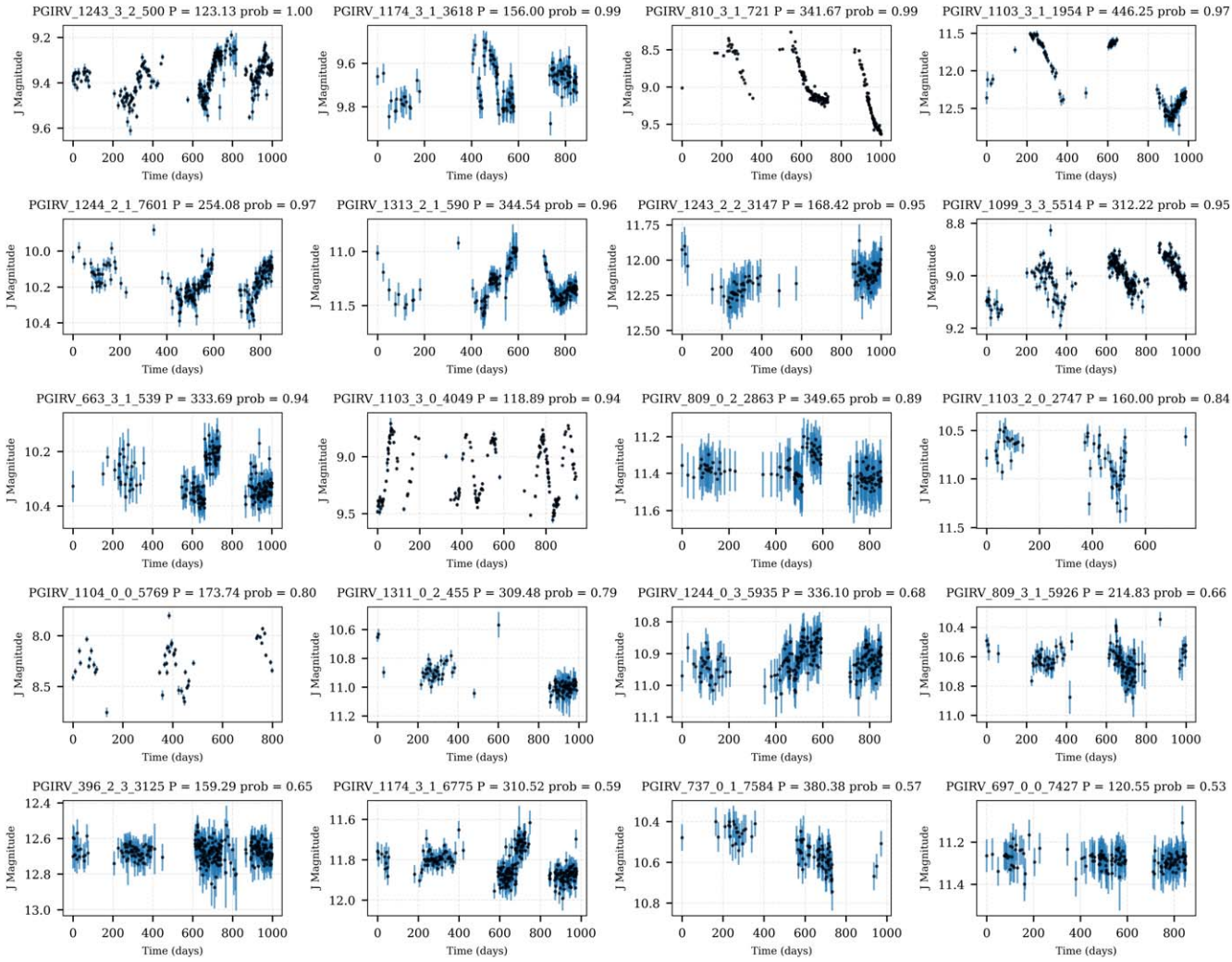

Figure 9. Example lightcurves of 20 LPVs, with the periods and LPV probabilites from our classifier. We see a good mix of large amplitude and small amplitude LPVs, showing a variety of morphologies. The period and internal designation of each LPV is listed. Lightcurves for all LPVs from the catalog are publicly available. A few suspicious lightcurves can also be seen in the above sample, which have a low classifier confidence (typically less than 0.6).

Table 2
PGIR Long Period Variables

Name	R.A.	Decl.	num_dets	lcdur	amplitude	bestperiod	J_mag	prob_lpv_sum
PGIRV_1008_0_3_357	172.2388	−7.2608	83	962.62	0.08	131	9.37	0.87
PGIRV_1012_0_1_365	190.4616	−7.1694	64	915.64	0.15	378	12.76	0.74
PGIRV_1012_2_0_355	192.9045	−8.4101	68	915.64	0.19	409	11.09	0.50
PGIRV_1015_0_2_1656	205.4475	−8.2296	69	892.62	0.12	155	12.62	0.97
PGIRV_1017_3_2_1044	218.1198	−5.5216	51	885.68	1.27	277	13.44	0.93
PGIRV_1018_1_3_1424	220.2763	−4.9539	66	904.68	0.06	240	11.33	0.53
PGIRV_1049_2_1_115	13.3383	−11.9760	90	989.17	0.21	195	11.29	0.99
PGIRV_1054_0_1_194	35.3839	−12.1281	96	838.81	0.07	173	10.04	0.71
PGIRV_1084_1_3_1287	183.5706	−9.0137	83	934.66	0.61	160	12.87	0.92
PGIRV_1086_0_3_472	194.2376	−11.7719	84	943.63	0.21	236	10.72	0.94
PGIRV_1091_3_1_1781	219.4370	−9.5846	73	909.66	0.07	230	10.13	0.62
PGIRV_1119_1_0_0	354.8956	−10.0279	71	975.25	0.08	188	10.19	0.53
PGIRV_1119_3_1_1517	357.4652	−9.8856	90	991.18	0.15	354	11.45	0.51
PGIRV_1120_2_2_851	4.0667	−17.6033	85	981.21	0.59	133	11.37	0.88
PGIRV_1121_3_0_1348	7.5163	−15.7706	17	642.21	1.06	103	14.49	0.54

Table 2
(Continued)

Name	R.A.	Decl.	num_dets	lcdur	amplitude	bestperiod	J_mag	prob_lpv_sum
PGIRV_1122_3_1_364	13.3731	−14.5094	88	989.22	0.27	175	10.77	0.99
PGIRV_1123_0_0_73	15.9838	−17.5140	97	993.18	0.15	159	11.01	0.94
PGIRV_1125_0_3_507	26.9506	−16.7215	42	784.83	0.22	242	13.87	0.55
PGIRV_1130_2_0_1225	52.6496	−17.9414	88	846.75	0.16	104	12.69	0.69
PGIRV_1157_1_1_652	185.4686	−13.8861	81	893.62	0.70	222	8.14	0.93
PGIRV_1190_1_0_597	350.5461	−15.3044	86	989.24	0.38	160	10.25	0.90
PGIRV_1191_1_1_948	355.1287	−13.8730	85	970.18	0.05	182	9.78	0.58
PGIRV_1192_2_0_573	3.1427	−22.9213	89	988.24	0.78	170	10.15	0.98
PGIRV_1198_1_3_1166	32.1860	−19.2157	10	751.92	0.81	257	14.43	0.55
PGIRV_1199_1_3_621	37.8213	−19.1320	104	838.77	0.19	237	10.24	0.98
PGIRV_1199_2_3_1245	39.9799	−21.0724	10	364.01	0.43	126	14.72	0.58
PGIRV_1256_2_3_445	333.9064	−22.0151	13	994.26	1.10	361	14.56	0.85
PGIRV_1258_1_2_436	341.5088	−20.4761	76	991.29	0.37	135	12.45	0.69
PGIRV_1260_1_1_910	349.8980	−18.9398	78	974.26	0.81	152	11.74	0.95
PGIRV_1260_2_1_184	353.3139	−21.5253	82	974.26	0.24	115	11.12	0.75
PGIRV_1263_3_3_824	9.7031	−23.9265	76	786.87	0.59	177	10.99	0.99
PGIRV_1266_2_3_1063	25.3007	−26.2789	41	788.87	0.24	232	13.90	0.59
PGIRV_126_1_1_1490	179.6802	63.3541	144	977.67	0.25	139	10.52	0.65
PGIRV_1296_1_0_1458	180.1042	−24.6867	29	837.74	0.60	175	9.82	0.59
PGIRV_1296_1_1_58	181.1569	−24.1606	80	876.65	0.27	139	11.17	0.91
PGIRV_1299_1_1_373	196.7535	−24.0919	73	877.65	0.46	162	9.07	0.99
PGIRV_129_1_1_808	209.4709	64.1117	141	930.77	0.69	138	11.57	0.86
PGIRV_130_2_2_1558	225.5413	59.5225	145	932.73	0.33	374	8.84	0.72
PGIRV_1313_0_3_2444	271.7071	−26.5023	34	735.93	0.25	150	10.92	0.87
PGIRV_1324_1_3_418	330.4792	−23.9745	66	994.25	0.07	163	9.34	0.74
PGIRV_1325_0_1_761	334.2913	−26.1176	68	987.29	0.88	341	11.57	0.93
PGIRV_1325_3_3_1070	337.9381	−23.5460	69	987.29	0.75	209	10.98	0.92
PGIRV_1326_0_1_412	339.7141	−26.9350	60	988.26	0.32	131	11.33	0.92
PGIRV_1327_1_1_724	344.6067	−24.1717	69	982.26	0.11	176	11.33	0.82
PGIRV_1329_1_0_161	355.8306	−25.2339	65	994.24	0.09	176	9.38	0.87

Note. Description of columns: *Name*: PGIR name for variable; *R.A.* and *decl.*: R.A. and decl. (in degrees for equinox 2000); *num_detections*: Number of detections; *lcdur*: Observation baseline; *amplitude* and *bestperiod*: amplitude and period; *J_mag*: mean *J* band magnitude; *prob_lpv_sum*: Classifier confidence (obtained by summing probability of being LPV and type II LPV; The complete catalog is available as a machine-readable table at Zenodo.

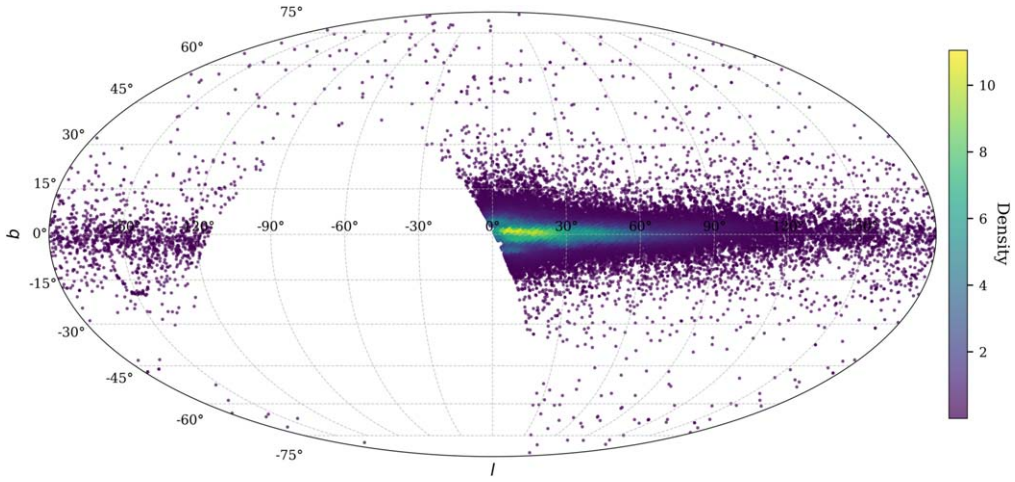

Figure 10. Spatial distribution of long-period variables from Gattini. A large fraction (150,414) of LPVs from the catalog lie within the galactic latitudes of −10° to +10°. The missing patch of sky is below a decl. of −28°, which PGIR cannot observe. The color bar represents the density of stars, with a high density around the galactic center.

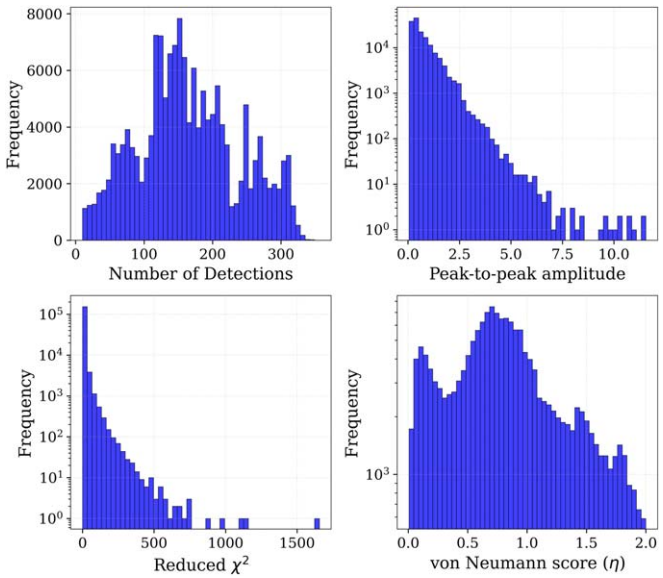

Figure 11. Distribution of select features in the LPV catalog. The majority of LPVs in the catalog (114,000) have amplitude less than 0.25 mag and are likely to be semiregular variables. The von Neumann score shows a bimodal distribution with peaks corresponding to Miras and semiregular variables.

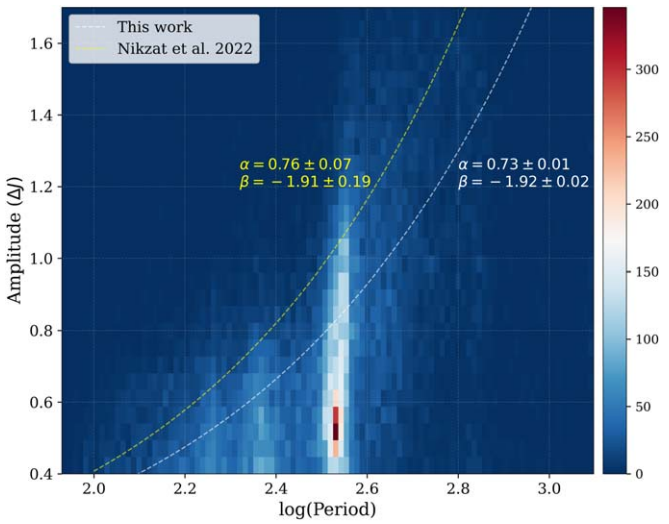

Figure 13. Period Amplitude relationship of Mira variables, chosen to be LPVs from the catalog having amplitude greater than 0.4 mag. The white curve represents the best-fit parameters from PGIR Miras, compared to the relation obtained by Nikzat et al. (2022). While the general trend of the period–amplitude relation holds, there is a significant scatter around the best-fit curve. The distribution of PGIR Miras in the log(period)–amplitude space is shown as a 2D histogram with the colorbar representing the number of objects in each bin.

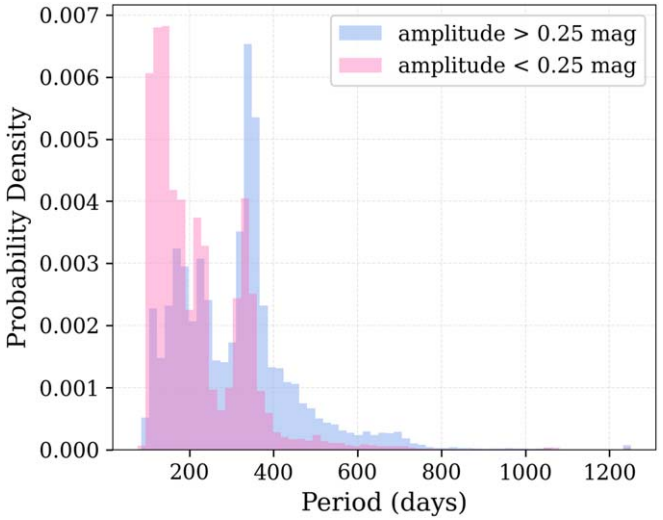

Figure 12. Distribution of periods in the PGIR LPV catalog, separated by amplitude, representing Miras (blue histogram) and semiregular variables (pink histogram). We show the density in each bin. We observe a bimodal period distribution for Miras with peaks at ∼200 and ∼340 days, while the semiregular variables show peaks at ∼130, ∼220, and ∼330 days.

4.1. Catalog Features

Figure 10 shows the spatial distribution of LPVs. We observe that 150,414 of LPVs are located between galactic latitudes (b_{gal}) of $-10°$ to $+10°$. We observe that a large number of LPVs have low amplitudes (<0.25) mag and a low percentile difference, implying that they are likely to be semiregular variables, while the LPVs showing amplitudes greater than 0.5 mag are likely to be Miras. Figure 11 shows the distribution of the peak-to-peak amplitude, the number of detections for each variable, the von Neumann score (η), and the reduced χ^2 of a sinusoid fit to a lightcurve.

The period histogram of LPVs is shown in Figure 12, with the Miras showing a doubly peaked distribution and the semiregular variables showing a triply peaked distribution. We obtain peaks at ∼200 and ∼340 days for the Miras, consistent with other studies (Glass et al. 2001; Matsunaga et al. 2005, 2009; Nikzat et al. 2022), which also find bimodal period distributions for Miras. The semiregular variables show peaks in their period distribution at ∼130, ∼220, and ∼330 days. Conroy et al. (2015) introduced a period–amplitude relationship for Miras:

$$\log_{10}\Delta J = \alpha \log_{10} P + \beta \tag{11}$$

We examine this relation for the Mira variables from PGIR in the *J*-band as shown in Figure 13 and compare it to the relation obtained in Nikzat et al. (2022). While the general trend holds, there is a significant scatter around the best-fit curve. We obtain $\alpha = +0.727 \pm 0.007$ and $\beta = -1.923 \pm 0.019$ from Miras in PGIR data. Our best-fit parameters agree reasonably with Nikzat et al. (2022), with $\alpha = +0.76 \pm 0.07$ and $\beta = -1.91 \pm 0.19$.

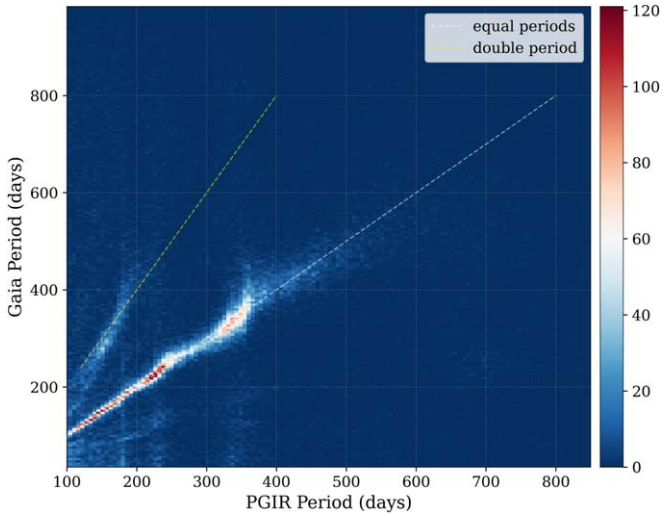

Figure 14. Comparison between periods of crossmatched LPVs from PGIR and Gaia. We observe a strong correlation between the two data sets. We also notice a secondary branch, which approximately represents the line where the period from Gaia is approximately twice the period from PGIR. The color gradient represents the number of objects in each bin of the 2D histogram.

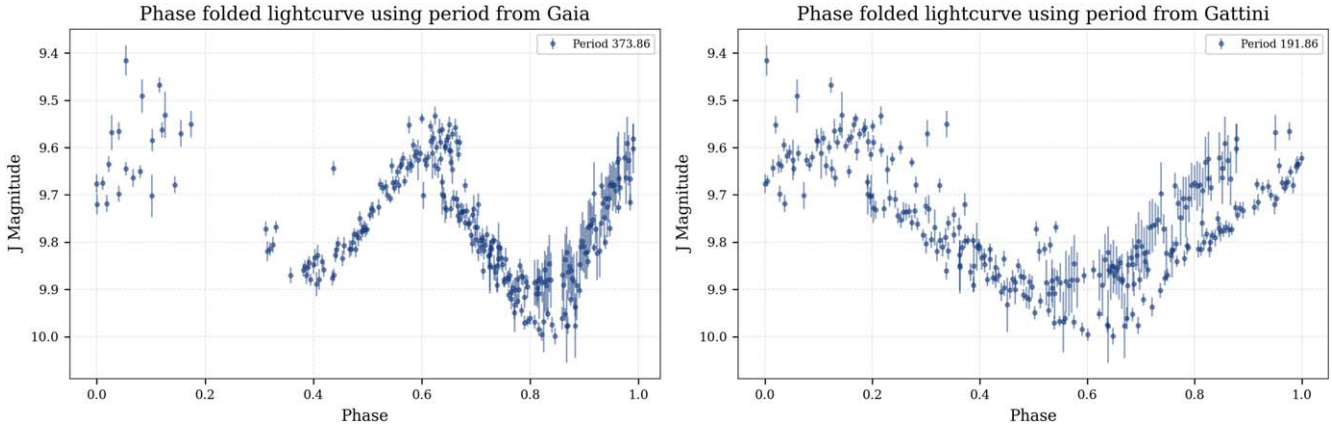

Figure 15. Example of an LPV where period estimated from Gattini lightcurve is more accurate than the one calculated from Gaia data.

4.2. Catalog Validation

The second catalog of long-period variables (Lebzelter et al. 2023) from Gaia data release 3 is the largest database of LPVs to date, with 1,720,588 entries spanning 34 months of Gaia data, with 391,914 LPVs having period information. We validate our catalog by comparing it with 150,195 LPVs from Gaia, above a decl. of $-28°$, that have calculated periods. We query the Gaia archive to retrieve the positions, periods, colors, and distances (whenever available) for these LPVs. We crossmatch the Gattini LPVs with Gaia keeping a maximum offset of $6''$ (corresponding to 2 PGIR pixels), to obtain 37,179 sources. Figure 14 shows the comparison between periods obtained from Gattini and Gaia. We observe a strong positive correlation with some scatter around the line of equal periods. We also observe a branch of correlated periods, where the Gaia period is approximately twice as much as the Gattini period, like the example shown in Figure 15. Visual inspection of these sources shows that Gaia overestimates the period of these variables, likely due to sparser coverage than PGIR.

Additionally, we also crossmatch our catalog with SIMBAD to understand the classification accuracy of the ML model. The SIMBAD website provides a collection of almost all variables previously found in the Catalina, OGLE, Gaia, and ASAS surveys; among others. We consider our LPV classification to be correct for a star if its previously assigned category from other surveys falls into one of "LongPeriodV*," "AGB*," "C*," "Mira," "LongPeriodV*_Candidate," "C*_Candidate," "Variable*," "Star," "OH/IR*," "MidIR," "NearIR," "Mira_Candidate" or "RGB*." Using a maximum crossmatch radius of $6''$,

Table 3
Simbad Types of Objects in the PGIR LPV Catalog

Category	Counts
New	70,369
LongPeriodV*_Candidate	39,272
LongPeriodV*	21,892
Mira	17,699
Star	1734
C*	794
Variable*	399
AGB*	267
OH/IR*	219
C*_Candidate	122
RGB*	71
MidIR	68
NearIR	56
Mira_Candidate	2
Others	1791

Note. Objects with confirmed sub-types are counted into their respective sub-types (e.g., Mira) and excluded from the broader class (e.g., LongPeriodV*).

we see that 70,369 (45.47% of the total LPVs) objects in the PGIR catalog are new. Of the remaining 84,386 objects, 1791 fall into categories other than the ones mentioned before, indicating a high purity (97.87%) of the PGIR catalog. The populations of the different Simbad categories are listed in Table 3.

5. Conclusion

We present a catalog of 159,696 LPVs from the PGIR survey created using a ML classifier. We assemble a training set of 4300 objects spanning LPVs showing clean sinusoidal evolution in their lightcurves, "Type II LPVs" showing non-sinusoidal but periodic lightcurves, non-LPVs comprised of R Cor Bor stars, Young Stellar Objects and other classes that display long timescale erratic variability, and non-variables. We train a gradient-boosted decision tree classifier on an artificially resampled data set using ADASYN upsampling and allKNN downsampling to class-balance the training set, achieving a true positive rate of 94.89% for the combined set of LPVs and Type II LPVs and an overall weighted g-mean score of 0.95, indicating high accuracy.

Following a feature extraction process for 35 million lightcurves from the PGIR survey at a rate of 0.1 s per lightcurve, we generate the LPV catalog using classifier predictions. We validate the catalog by crossmatching it with the Gaia catalog of LPVs from the third data release (Lebzelter et al. 2023) and the Simbad interface to ascertain purity and completeness. We find that 73,346 variables are newly discovered, which is 45.9% of the total catalog size, and find that 1,867 objects in the catalog fall out of desired Simbad classification categories, indicating a purity of ∼97.8%. We also find a good correlation in periods between LPVs from PGIR and Gaia. We use the PGIR LPV catalog to examine the period–amplitude relationship for Miras in the J band and verify the bimodality in period distributions of Miras found in other studies (Glass et al. 2001; Matsunaga et al. 2005, 2009; Nikzat et al. 2022). The methods developed in this paper will be useful for more sensitive searches for periodic variables with the Vera Rubin Observatory (Ivezić et al. 2019).

Acknowledgments

We thank the referee for useful suggestions that helped improve this manuscript. Palomar Gattini-IR (PGIR) is generously funded by Caltech, Australian National University, the Mt Cuba Foundation, the Heising Simons Foundation, the Binational Science Foundation. PGIR is a collaborative project among Caltech, Australian National University, University of New South Wales, Columbia University and the Weizmann Institute of Science. M.M.K. acknowledges generous support from the David and Lucille Packard Foundation. M.M.K. acknowledges the US-Israel Bi-national Science Foundation grant 2016227. M.M.K. acknowledges the Heising-Simons foundation for support via a Scialog fellowship of the Research Corporation. M.M.K. and A.M.M. acknowledge the Mt Cuba foundation. J.S. is supported by an Australian Government Research Training Program (RTP) Scholarship. A.S. acknowledges support from the Caltech Summer Undergraduate Research Fellowship (SURF) funded by the NSF Astronomy and Astrophysics grant No. 2206730. K.D. was supported by NASA through the NASA Hubble Fellowship grant #HST-HF2-51477.001 awarded by the Space Telescope Science Institute, which is operated by the Association of Universities for Research in Astronomy, Inc., for NASA, under contract NAS5-26555. R.S. acknowledges grant No. 12073029 from the National Natural Science Foundation of China (NSFC).

ORCID iDs

Aswin Suresh https://orcid.org/0009-0005-8230-030X
Viraj Karambelkar https://orcid.org/0000-0003-2758-159X
Mansi M. Kasliwal https://orcid.org/0000-0002-5619-4938
Michael C. B. Ashley https://orcid.org/0000-0003-1412-2028
Kishalay De https://orcid.org/0000-0002-8989-0542
Matthew J. Hankins https://orcid.org/0000-0001-9315-8437
Jamie Soon https://orcid.org/0000-0001-9226-4043
Roberto Soria https://orcid.org/0000-0002-4622-796X
Tony Travouillon https://orcid.org/0000-0001-9304-6718

References

Alcock, C., Allsman, R. A., Alves, D. R., et al. 2001, ApJ, 554, 298
Ambikasaran, S., Foreman-Mackey, D., Greengard, L., Hogg, D. W., & O'Neil, M. 2015, ITPAM, 38, 252
Bellm, E. C., Kulkarni, S. R., Graham, M. J., et al. 2019, PASP, 131, 018002

Bergstra, J., & Bengio, Y. 2012, J. Mach. Learn. Res., 13, 281
Boone, K. 2019, AJ, 158, 257
Breiman, L. 2001, Mach. Learn., 45, 5
Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. 1984, Classification and Regression Trees (1st ed.; New York: Chapman and Hall/CRC), 368, https://api.semanticscholar.org/CorpusID:29458883
Chen, X., Wang, S., Deng, L., de Grijs, R., & Yang, M. 2018, ApJS, 237, 28
Chen, X., Wang, S., Deng, L., et al. 2020, ApJS, 249, 18
Christy, C. T., Jayasinghe, T., Stanek, K. Z., et al. 2023, MNRAS, 519, 5271
Clayton, G. C. 2012, JAAVSO, 40, 539
Conroy, C., van Dokkum, P. G., & Choi, J. 2015, Natur, 527, 488
Cutri, R. M., Skrutskie, M. F., van Dyk, S., et al. 2003, yCat, II/246
Cutri, R. M., Wright, E. L., Conrow, T., et al. 2012, yCat, II/311
De, K., Hankins, M. J., Kasliwal, M. M., et al. 2020, PASP, 132, 025001
Debosscher, J., Sarro, L. M., Aerts, C., et al. 2007, A&A, 475, 1159
Dekany, I., Catelan, M., Minniti, D. & the VVV Collaboration 2011, arXiv:1111.0909
Foreman-Mackey, D., Agol, E., Ambikasaran, S., & Angus, R. 2017, AJ, 154, 220
Glass, I. S., Matsumoto, S., Carter, B. S., & Sekiguchi, K. 2001, MNRAS, 321, 77
He, H., Bai, Y., Garcia, E. A., & Li, S. 2008, in 2008 IEEE Int. Joint Conf. Neural Networks (IEEE World Congress on Computational Intelligence) (Hong Kong: IEEE), 1322
Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111
Jayasinghe, T., Stanek, K. Z., Kochanek, C. S., et al. 2019, MNRAS, 486, 1907
Karambelkar, V. R., Adams, S. M., Whitelock, P. A., et al. 2019, ApJ, 877, 110
Karambelkar, V. R., Kasliwal, M. M., Tisserand, P., et al. 2021, ApJ, 910, 132
Kasliwal, M., Jencson, J., Lau, R., et al. 2018, SPIRITS: SPitzer InfraRed Intensive Transients Survey, Spitzer Proposal ID #14089
Ke, G., Meng, Q., Finley, T., et al. 2017, LightGBM: a highly efficient gradient boosting decision tree, in Advances in Neural Information Processing Systems, Vol. 30 *(Long Beach, California, USA)* ed. I. Guyon et al. (Red Hook, NY: Curran Associates, Inc.), 3149
Kim, D.-W., & Bailer-Jones, C. A. L. 2016, A&A, 587, A18
Lebzelter, T., Mowlavi, N., Lecoeur-Taibi, I., et al. 2023, A&A, 674, A15
Lemaître, G., Nogueira, F., & Aridas, C. K. 2017, J. Mach. Learn. Res., 18, 1
Lomb, N. R. 1976, Ap&SS, 39, 447
Masci, F. J., Hoffman, D. I., Grillmair, C. J., & Cutri, R. M. 2014, AJ, 148, 21
Masci, F. J., Laher, R. R., Rebbapragada, U. D., et al. 2017, PASP, 129, 014002
Matsunaga, N., Fukushi, H., & Nakada, Y. 2005, MNRAS, 364, 117
Matsunaga, N., Kawadu, T., Nishiyama, S., et al. 2009, MNRAS, 399, 1709
Nikzat, F., Ferreira Lopes, C. E., Catelan, M., et al. 2022, A&A, 660, A35
Pojmanski, G. 2002, AcA, 52, 397
Pojmanski, G., Pilecki, B., & Szczygiel, D. 2005, AcA, 55, 275
Qu, H., Sako, M., Möller, A., & Doux, C. 2021, AJ, 162, 67
Samus', N. N., Kazarovets, E. V., Durlevich, O. V., Kireeva, N. N., & Pastukhova, E. N. 2017, ARep, 61, 80
Scargle, J. D. 1982, ApJ, 263, 835
Shappee, B. J., Prieto, J. L., Grupe, D., et al. 2014, ApJ, 788, 48
Skrutskie, M. F., Cutri, R. M., Stiening, R., et al. 2006, AJ, 131, 1163
Smak, J. I. 1966, ARA&A, 4, 19
Soszyński, I., Udalski, A., Szymański, M. K., et al. 2009, AcA, 59, 239
Stetson, P. B. 1996, PASP, 108, 851
Sánchez-Sáez, P., Reyes, I., Valenzuela, C., et al. 2021, AJ, 161, 141
Tomek, I. 1976, ITSMC, SMC-6, 448
Udalski, A., Szymanski, M., Kaluzny, J., Kubiak, M., & Mateo, M. 1992, AcA, 42, 253
VanderPlas, J. T., & Ivezić, Ž. 2015, ApJ, 812, 18
von Neumann, J. 1941, The Annals of Mathematical Statistics, 12, 367
Williams, C., & Rasmussen, C. 1995, Gaussian processes for regression, in Advances in Neural Information Processing Systems, Vol. 8 *(Denver, Colorado)* ed. D. Touretzky, M. Mozer, & M. Hasselmo (Cambridge, MA: MIT Press), 514
Wood, P. R., Alcock, C., Allsman, R. A., et al. 1999, in IAU Symp. 191, Asymptotic Giant Branch Stars, ed. T. Le Bertre, A. Lebre, & C. Waelkens, 151
Wray, J. J., Eyer, L., & Paczyński, B. 2004, MNRAS, 349, 1059
Wright, E. L., Eisenhardt, P. R. M., Mainzer, A. K., et al. 2010, AJ, 140, 1868