The Astrophysical Journal Supplement Series, 203:32 (27pp), 2012 December doi:10.1088/0067-0049/203/2/32

CONSTRUCTION OF A CALIBRATED PROBABILISTIC CLASSIFICATION CATALOG: APPLICATION TO 50k VARIABLE SOURCES IN THE ALL-SKY AUTOMATED SURVEY

Joseph W. Richards[1,2], Dan L. Starr[1], Adam A. Miller[1], Joshua S. Bloom[1],
Nathaniel R. Butler[3], Henrik Brink[1], and Arien Crellin-Quick[1]
[1] Astronomy Department, University of California, Berkeley, CA 94720-3411, USA; jwrichar@stat.berkeley.edu
[2] Statistics Department, University of California, Berkeley, CA 94720-7450, USA
[3] School of Earth and Space Exploration, Arizona State University, Tempe, AZ 85287, USA

Received 2012 April 19; accepted 2012 September 8; published 2012 December 3

ABSTRACT

With growing data volumes from synoptic surveys, astronomers necessarily must become more abstracted from the discovery and introspection processes. Given the scarcity of follow-up resources, there is a particularly sharp onus on the frameworks that replace these human roles to provide accurate and well-calibrated probabilistic classification catalogs. Such catalogs inform the subsequent follow-up, allowing consumers to optimize the selection of specific sources for further study and permitting rigorous treatment of classification purities and efficiencies for population studies. Here, we describe a process to produce a probabilistic classification catalog of variability with machine learning from a multi-epoch photometric survey. In addition to producing accurate classifications, we show how to estimate *calibrated* class probabilities and motivate the importance of probability calibration. We also introduce a methodology for feature-based anomaly detection, which allows discovery of objects in the survey that do not fit within the predefined class taxonomy. Finally, we apply these methods to sources observed by the All-Sky Automated Survey (ASAS), and release the Machine-learned ASAS Classification Catalog (MACC), a 28 class probabilistic classification catalog of 50,124 ASAS sources in the ASAS Catalog of Variable Stars. We estimate that MACC achieves a sub-20% classification error rate and demonstrate that the class posterior probabilities are reasonably calibrated. MACC classifications compare favorably to the classifications of several previous domain-specific ASAS papers and to the ASAS Catalog of Variable Stars, which had classified only 24% of those sources into one of 12 science classes.

Key words: catalogs – methods: data analysis – methods: statistical – stars: variables: general – techniques: photometric

Online-only material: color figures, machine-readable table

1. INTRODUCTION

Synoptic imaging surveys have begun to routinely collect dozens to thousands of epochs of photometric data over wide swaths of the sky. The manifest destiny for optical time-domain studies is the Large Synoptic Survey Telescope (LSST; Tyson 2002), which will collect time histories for $\mathcal{O}(10^9)$ stars and explosive transients. With data collected for so many sources, no longer is it possible for experts to manually scrutinize significant subsets of the data. In this era of wide-field time-domain surveys, accurate multi-class source catalogs, which are created automatically by machine-learning algorithms, can greatly help maximize the scientific output from these projects (Eyer et al. 2008; Borne et al. 2009). Furthermore, with imperfect and limited information on each source, variability and transient catalogs must be probabilistic in nature, with well-calibrated posterior class probabilities. This enables each scientist to use a personalized threshold for selecting objects for follow-up, where science class probabilities fit naturally within a framework for optimizing the allocation of limited resources, and to select objects for population studies, where a rigorous treatment requires detailed understanding of the purity and efficiency of the sample.[4]

Creating probabilistic multi-class catalogs for large-scale time-domain photometric surveys is a difficult task. First and foremost, a set of salient class-predictive features[5] needs to be estimated for each source. From unevenly sampled light curves that contain seasonal gaps, varying levels of noise, and occasional spurious flux measurements, estimating the periods and amplitudes of oscillations for each source is not trivial. Furthermore, devising light-curve features that can separate specific subclasses of sources requires deep domain knowledge. Next, classification models must be constructed to map the light-curve feature vector for each source to a set of posterior class probabilities. These classifiers need to be able to automatically learn multiple class boundaries in high-dimensional feature space from a set of training data with known classes and, for each source, return a calibrated posterior class probability for each science class. This endeavor is complicated by the fact that the set of training data is typically not representative of the objects in the survey, which can cause large sample-selection biases (see Richards et al. 2012) in the posterior class probability estimates. Additionally, the sources observed by the survey are not guaranteed to fit neatly into any of the predefined classes, necessitating anomaly detection to identify which sources are likely to belong to an undefined science class.

Several aspects of the autonomous cataloging effort have required focused research attention. In Richards et al. (2011),

[4] Here we define the classification purity of a particular class as one minus the ratio of the number of objects falsely classified as belonging to that class to the total number of objects classified into the class (i.e., one minus false positive rate). Likewise, the classification efficiency is the ratio of the number of objects correctly classified into that class to the total number of objects in the class (i.e., one minus the false negative rate).

[5] We define a feature to be a deterministic real-numbered or categorical metric based on the time-series input or spatial location of the source. See Section 3.

we introduced an end-to-end framework for machine-learned classification of variable stars, with advancements in periodic and non-periodic light-curve feature estimation as well as probabilistic, non-parametric classification methodology. In terms of classification error rate, our methods showed significant improvement over the previous state of the art (Debosscher et al. 2007) on a well-studied data set. Indeed, other groups have also converged onto a similar set of tools as the best current light-curve classification methodology for variable stars (e.g., Debosscher et al. 2009; Dubath et al. 2011; Blomme et al. 2011; Varón et al. 2011). In Richards et al. (2012), we introduced a methodology to overcome the debilitating effects of non-representative training sets on variable star classification, and in Long et al. (2012) we devised methods to appropriately use light-curve data from older surveys to classify periodic variable stars in new surveys. With these advances, the accuracy of variable star classification is improving demonstrably, with cross-validated error rates approaching 15%–20% on multi-class problems with different data sets (Dubath et al. 2011; Richards et al. 2011).

In this paper, we build on these recent advancements in the photometric classification of variability by focusing on the problem of *how to properly construct a variable star classification catalog from a photometric survey*. Accurate classification of each source in the survey remains the primary goal of this endeavor. However, there are several other issues that arise when generating classification catalogs for use in astrophysical studies. First, a classification catalog requires good calibration for the posterior class probability estimates, $\mathbf{P}(\text{class}|\text{survey data})$. Good calibration means that of all the objects for which we estimate a posterior class probability, p, of belonging to a certain science class, p proportion of them *truly* belong to that class. In this paper, we describe a method for calibrating classifier probabilities and outline how such information can and should be used when employing such a probabilistic classification catalog for downstream astrophysical inference.

Second, when constructing a classification catalog for a large number of objects, anomalies will certainly be present. When building a supervised classification model, these anomalies are typically not accounted for, resulting in a classification schema which attempts to artificially coerce each object into a predefined classification taxonomy. In this paper, we describe the use of a semi-supervised anomaly detection routine which allows a determination of which sources do not resemble any of the training data and likely belong to a variability class not populated by the training set. We determine, for each source in the catalog, a real-valued measure of the relative degree of deviance of that source from the training data.

Third, in a photometric survey, each object may also have cataloged information in other passbands. We detail, in this paper, how to probabilistically associate sources with objects in external catalogs based on positional information as well as photometric features. This allows us to obtain further classification features (e.g., color) which are crucial for allowing accurate classification statements. In addition, we use a non-parametric method to impute the values of those attributes when no match is detected, allowing us to use any classification method that requires complete data.

Finally, we use this methodology to create a calibrated probabilistic classification catalog for a set of 50,124 sources in the All-Sky Automated Survey (ASAS; Pojmański 1997) based on its publicly available ASAS V-band light curve and colors. Our Machine-learned ASAS Classification Catalog (MACC, publicly available online) contains, for each source, posterior probabilities for 28 different science classes. This is a wealth of new information compared to the existing ASAS Catalog of Variable Stars (ACVS; Pojmański 2002), which classified a subset of these sources into 12 science classes without supplying any posterior class probabilities and giving the uninformative class label "MISC" to a majority of objects. In addition to probabilistic classifications, MACC gives an anomaly score for each ASAS source, which describes its proximity to objects in the training set. Furthermore, our catalog provides updated periods, peak-to-peak amplitudes, and dozens of other estimated features for each ASAS light curve. We ensure that all steps in the MACC catalog creation are transparent and provide a public interface to the catalog at www.bigmacc.info.

2. DATA

2.1. ASAS Data Collection

The All-Sky Automated Survey,[6] is an ongoing, long-term project dedicated to the detection and monitoring of the photometric variability of bright stars (Pojmański 1997). Since 2000 August, ASAS has monitored bright stars ($V < 14$ mag) in the entire available sky south of $\delta < +28^\circ$ from Las Campanas Observatory. ASAS uses two small wide-field telescopes to monitor the sky with V- and I-band filters. Each ASAS telescope takes repeated 180 s exposures using a 2 K × 2 K CCD camera with 15 μm pixels, covering 8.5×8.5 deg^2 of the sky (see Pojmański 1997 for further details).

To date, ASAS has taken more than 267,260 V-band frames, imaging approximately 17 million stars of V-band magnitude between 8 and 14. Of these 17 million objects, ASAS has identified 50,124 variable stars and published the results in the ACVS (Pojmański 2002, 2003; Pojmański & Maciejewski 2004, 2005; Pojmański et al. 2005). The catalog, which contains a rough classification for each source, is made publicly available through the ASAS Web site, along with V-band light curves for 15 million ASAS sources. For the 50,124 variable stars, ASAS has retrieved a median of 541 usable epochs of V-band measurements. The ACVS light curves range in the number of good detection epochs from 3 to 2232.

2.2. ASAS Photometric Light Curves

We retrieved the ASAS ACVS data set by first referencing the ACVS.1.1 catalog, which contains 50,124 variable stars, and individually retrieving the data for each source from the ACVS Web site.[7] These sources were imported into our DotAstro.org (http://dotastro.org) astronomical light-curve warehouse for visualization and use with internal frameworks (Brewer et al. 2009). Each ASAS source's time-series data file is partitioned by its observed field and includes information on the quality of the aperture photometry for each epoch, as well as magnitude measurements (and uncertainties) from up to five different apertures. Prior to importing the data we chose a single aperture for each epoch using the method detailed below and excluded epochs with a quality `GRADE=D` or quality `GRADE=C` when `MAG=29.999`, which, as detailed in the ASAS data files, corresponds to a non-detection. Given to the undue influence of extreme photometric outliers in light-curve feature estimation, before the generation of time-series based features, we applied

[6] http://www.astrouw.edu.pl/asas/

[7] Available at http://www.astrouw.edu.pl/asas/?page=acvs.

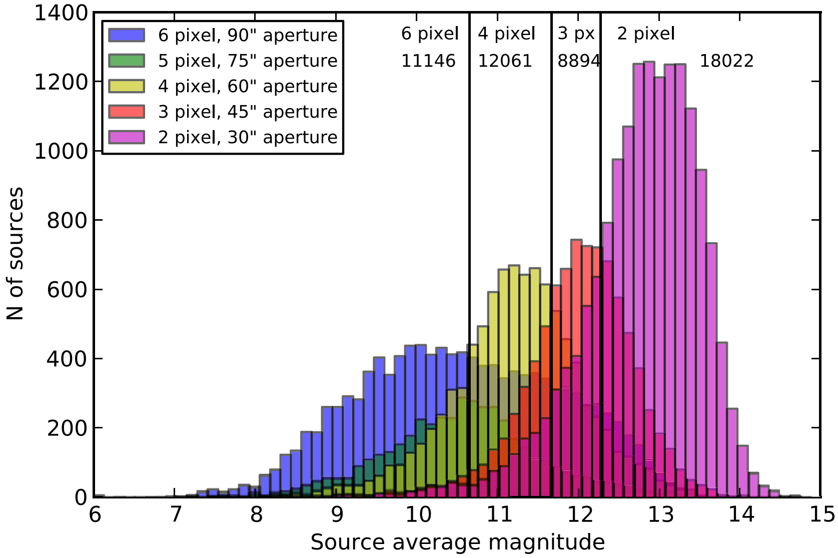


Figure 1. Aperture-wise histograms of the average magnitude of ASAS sources whose minimal magnitude dispersion was observed to be in the specified aperture. As expected, brighter sources experience smaller dispersion when observed in wider apertures and fainter sources show smaller dispersion in narrower apertures. Using these histograms, we construct a kernel density estimation classifier to determine the optimal aperture to use for each ASAS source as a function of its average magnitude of brightness. The magnitude cuts from this procedure are overlaid in vertical lines and the total number of ASAS objects extracted with each aperture is listed in the figure.

(A color version of this figure is available in the online journal.)

sigma clipping to each ASAS light curve, excluding observations that lie beyond four standard deviations from each source's mean magnitude.

ASAS provides five aperture measurements using annuli ranging from 2 pixels (30″) to 6 pixels (90″). Although the ASAS team outlined a basic algorithm for choosing which aperture to use for each source given its average magnitude (Pojmański et al. 2005), we decided to use our own magnitude-dependent aperture cuts. Our procedure is the following: we begin by determining the aperture which has the minimum magnitude variance for a source.[8] The idea behind using minimum magnitude variance is that apertures that are too small will not capture all of the source's flux, resulting in larger Poisson noise in the measured brightness of the object, whereas apertures that are too large will incur more background noise and noise due to contamination from neighboring sources. For each aperture, we can visualize the distribution of mean magnitudes for the sources whose minimal magnitude dispersion occurred in that aperture (see Figure 1). This information was subsequently employed to construct a simple kernel density estimation (KDE) classifier to determine the optimal aperture to use for each source, as a function of its mean magnitude. Using this classifier we determine the optimal aperture for each ASAS source, as a function of its average magnitude (more precisely, the median magnitude of its five aperture-wise average magnitudes), and only import the light-curve measurements from that aperture. The optimal magnitude cuts for each aperture are overplotted in Figure 1.

We also assess the sensitivity of the features and classifications to the aperture size. Beginning with the optimal KDE-selected aperture size for each source, we perturbed the aperture size for all ASAS objects using (a) 1 pixel larger than the optimal aperture size and (b) 1 pixel smaller than the optimal. Objects whose optimal aperture size was initially the largest (for a) or smallest (for b) were left unchanged. Taking the feature sets generated in (a) and (b), we regenerated the variable star classification for 5000 arbitrary ASAS sources (10% of the full catalog). We found that 71.7% and 73.9% of all sources have the same classification as that of the optimal aperture, respectively, for experiments (a) and (b). Most of the differences occur for sources with low classification probabilities, as only (a) 43.2% and (b) 43.2% of objects with classification probability of less than 0.5 of belonging to the most likely class in the optimal-aperture case retain the same classification; this probability jumps to (a) 82.7% and (b) 85.7% for classifications of higher than 0.5 probability. This underscores the importance of rigorously choosing the optimal aperture for each light curve, as we have done with the KDE classifier, because faced with such high sensitivity of the classifications to aperture size, it is imperative that we classify the *best* light curve.

2.3. Obtaining Source Colors from a Machine-learned Cross-match

In addition to information gleaned from single-band light curves, color information is invaluable to classifying variable stars. To generate color features, we use the Naval Observatory Merged Astrometric Dataset (NOMAD; Zacharias et al. 2004) to obtain single-epoch B-, V-, R-, J-, H-, and K_s-band photometry for each ASAS object, which we use to compute five color features ($B - V$, $V - R$, $R - J$, $J - H$, $H - K_s$) for each source. Although the ASAS ACVS catalog provides cross-correlated Two Micron All Sky Survey (2MASS) colors, the additional optical filters provided by NOMAD supplies a richer set of colors to aid the classifier. Due to the large ASAS positional errors, we decided against using simple spatial cross-correlation to match each ASAS source to the NOMAD catalog. Instead, we train a machine-learned classifier which takes as input seven positional and photometric features to determine whether a NOMAD candidate is indeed a match to the ASAS star. In addition to the separation distance between the ASAS source and NOMAD candidate, we employ the NOMAD nearest neighbor rank (ordered by distance from the ASAS

[8] Within the field with the greatest number of observations for that source.

source), magnitude differences in J, H, and K_s bands and $J - K_s$ color difference between ACVS and *NOMAD*, and the V-band difference between the ASAS light-curve mean magnitude and NOMAD to allow a richer view of each source which will facilitate the ASAS–NOMAD matching procedure.

This ASAS–NOMAD association classifier was initially trained using 48 ASAS sources of known class, sampled from 24 science classes, with 2 sources from each science class taken from the literature. For each of these training objects, we manually determined which source, from a NOMAD catalog query around the position of the ASAS source, was truly associated with that object. These sources were classified as "match," while all other sources returned by the NOMAD query were classified "non-match."

Using the seven positional and photometric features, we initially trained a random forest (RF; see Breiman 2001) classifier on the 48-object training set and applied the classifier to predict match/non-match for a sample of 30,000 of the ~500,000 NOMAD sources which are retrieved when the NOMAD catalog is queried around each of the 50,124 ASAS sources in our data set. Using the active learning technique of Richards et al. (2012), in each iteration of learning we selected 17 NOMAD sources which would have high impact in improving the performance of the classifier, and manually classified each as a "match" or "non-match," and subsequently added these objects to the training set. This active learning process was performed over 10 iterations, resulting in a robust classifier which can accurately and automatically decide whether a NOMAD source is associated with an ASAS source based on the positional and photometric features.

Ultimately, the classification algorithm was applied to each ASAS source to find the matching NOMAD entry, if any. For each ASAS star, we find the NOMAD source with the highest classifier probability of "match," with a preference for spatially closer matches when identical probabilities are returned for multiple NOMAD sources. If, for an ASAS object, no NOMAD source achieves "match" probability >50%, then we decide that no NOMAD source exists for that object. When applied to all 50,124 ASAS sources, we find that 93.9% of these sources match a NOMAD source. Perturbing the NOMAD positions within their errors (which are provided by the catalog) and repeating the match procedure shows that only 2 of 675 sources of known association (match or non-match) are incorrectly classified by our procedure, giving us confidence that our cross-matching error rate is <1%.

For the 47,044 objects with a NOMAD match, we extract five NOMAD color features for use in the variable star classifier. For the remaining 3080 objects with no NOMAD match, we impute their colors using the MissForest imputation routine of Stekhoven & Bühlmann (2012). MissForest[9] is an imputation routine that uses a series of RFs to predict the value of each missing feature based on the observed features for that source. The MissForest algorithm builds an RF regression model (for real-valued features) or classifier (for categorical features) to predict the value of each feature from all of the other features. Beginning from some initialization of the missing features, the algorithm iterates until convergence is attained and outputs the predicted value for each missing feature in the data matrix. On multiple data sets, Stekhoven & Bühlmann (2012) show that MissForest outperforms other common methods, such as K-nearest neighbors and Lasso, in imputation accuracy. We employ MissForest using 100 trees.

Table 1
Color Imputation Median Absolute Errors using the MissForest Imputation Method

Color	σ
$B - J$	0.965
$H - K_s$	0.059
$J - H$	0.087
$R - J$	0.751
$V - J$	0.863

We test the accuracy of MissForest in imputing variable star colors by the following experiment. Starting with the set of 47,044 objects with a satisfactory NOMAD match, we null out the colors for a random 6.1% of the objects (the same fraction of ASAS objects with no NOMAD match). Then, using the leftover set of sources with known colors, we impute the nulled out colors using MissForest. This allows us to compare the true colors to the imputed colors for this subset of data, which we do using median absolute error (MAE),

$$\sigma(\mathbf{x}_{j,\mathrm{imp}}) = \mathrm{median}_i |x_{ij,\mathrm{true}} - x_{ij,\mathrm{imp}}|, \quad (1)$$

where $x_{ij,\mathrm{true}}$ and $x_{ij,\mathrm{imp}}$ denote the true and imputed values, respectively, of color j for object i. MAEs, σ, for each of the five colors in our data are reproduced in Table 1. While the MAEs for each color, particularly the optical–NIR colors, are larger than the typical uncertainty of the observed color for any individual source, we note that a large scatter is to be expected because we are imputing the observed color without reddening corrections. Indeed, an examination of the observed color for each class shows that the typical within-class scatter is $\gtrsim$2 mag, most likely owing to the various galactic latitudes at which the ASAS sources are observed. The imputation procedure confidently identifies stars as being either red or blue, and the obtained accuracy of these imputations is similar to the typical scatter in the observed colors, which gives us confidence that the procedure is sufficient for classification purposes.

3. ASAS VARIABLE STAR CLASSIFIER

Probabilistic supervised light-curve classification has recently received much attention in the literature. For example, Debosscher et al. (2007), Dubath et al. (2011), and Richards et al. (2011) have applied modern machine-learning methods to ~25-class variable star problems, using photometric light-curve data from the *Hipparcos* and OGLE surveys to classify sources into a fine taxonomy of periodic variables (e.g., Cepheids, RR Lyrae, and their subclasses, eclipsing binaries including detached, semi-detached and contact systems) and non-periodic variables (e.g., T Tauri and other young stellar objects, S Doradus hypergiants, and evolved stars such as Wolf-Rayet variables). This automated classification methodology consists of the following two-step process.

1. From each light curve, a set of m features (e.g., period, amplitude, etc.) is extracted. These features are constructed to capture the class-predictive information encoded within each light curve.
2. Using a training set of objects of known class, a classification model, which maps from m-dimensional feature space to the set of classes, is fit. Methods such as neural nets, decision trees, support vector machines, and RFs are

[9] The R package missForest is freely available at http://cran.r-project.org/web/packages/missForest.

classification models that have been used for light-curve classification. The fitted classification model serves as a class-prediction engine.

Once the classifier has been trained, it is trivial to predict the class of each variable star, which entails first extracting the feature vector of the object and subsequently inserting that vector into the classifier to obtain a prediction. Many classifiers, such as RF, produce a vector of posterior class probabilities for each object.

To construct the ASAS variable star classification catalog, we use a set of $m = 72$ features: 67 light-curve features and 5 colors (described in Section 2.3). See Section 3.1 below for a description of the features used. We use an RF classifier, which has been shown to attain high levels of accuracy in variable star classification by Dubath et al. (2011) and Richards et al. (2011). Richards et al. (2011) found that the RF classifier attained the lowest error rates in classifying *Hipparcos* and OGLE variable stars in a side-by-side comparison with a dozen other classification models. In Section 3.2 we describe how to attain a training set for ASAS in order to minimize classification errors due to sample-selection bias (see Richards et al. 2012 for a thorough discussion of sample-selection bias for light-curve classification).

3.1. Light-curve Feature Extraction

Raw light-curve data consist of measurements of a source's brightness over unevenly sampled epochs. From these data, our challenge is to estimate a set of features that are predictive of each source's class (e.g., it is well known that period, amplitude, and color are all highly predictive of class for certain classes of pulsating variable stars), while being agnostic to other latent factors that are unrelated to (or at most, mildly correlated with) an object's science class. Examples of such latent factors are that each ASAS light curve consists of a different number of epochs (ranging from 3 to 2232 epochs with median of 541), over a different time baseline, with distinct noise properties and differing cadences. Furthermore, each ASAS source has a unique mean brightness (from 4th to 15th magnitude in *V*), resides in a unique position in the sky, and has its light affected by more or less intervening dust.

We have constructed a set of 66 light-curve features meant to capture the essence of photometric variability of the science classes of interest, and have written algorithms that efficiently compute these features from light-curve data, in an average of 4.5 s per ASAS light curve. In Richards et al. (2011), a set of 52 features was used to represent each variable star. Below, we describe the additional features that have been used in this study, and also outline some modifications to the algorithms used for periodic modeling.

3.1.1. Computationally Efficient Regularized Fitting of Periodic Signals of Arbitrary Shape

In this study, we employ a novel fitting routine which seeks to simultaneously discover the true period of a source while also modeling the light curve in detail.

We begin by applying our fast Lomb–Scargle algorithm (fit of single sinusoid; Richards et al. 2011), which uses the method of Zechmeister & Kürster (2009) to discover all marginally significant periods for a given light curve on a broad frequency test grid ($\nu_{\rm min} = 1/T$, $\nu_{\rm max} = 10$, $\delta\nu = 0.1/T$ cycle/day, where T is the data timespan). For test frequencies where the power spectrum has a value > 6 (i.e., $<1\%$ of test points, corresponding to roughly 3.5σ significance), we fit a multi-harmonic model,

$$m_i = ct_i + \sum_{n=1}^{8} A_n \sin(2\pi \nu_0 n t_i) + B_n \cos(2\pi \nu_0 n t_i) + b_{n,o} \quad (2)$$

consisting of a sinusoid at the initial frequency, ν_0, plus sinusoids at each of the $n = (2, \ldots, 8)$ harmonics of that initial frequency and a constant offset, $b_{n,o}$, for each harmonic. We choose $n = 8$ to allow for sufficient model complexity to account for the light curves under study. The fitting of model 2 is performed with a regularization penalty to avoid overfitting, and the number of effective model degrees of freedom is typically well below the allowed value of $2 \times 8 = 16$.

In the fitting, we minimize

$$R = \sum_{i=1}^{N} \frac{(d_i - m_i)^2}{\sigma_i^2} + N\lambda \times \sum_{n=1}^{8} n^4 \left(A_n^2 + B_n^2\right), \quad (3)$$

with respect to the model parameters θ and the regularization parameter λ. Here, the photometric data are d_i, the model is m_i, N is the number of data points, and $\sqrt{A_n^2 + B_n^2}$ is the amplitude of the nth Fourier harmonic. The second term above effectively penalizes the model in proportion to the magnitude of its second derivative. Small values of λ result in models with high-frequency structure, whereas large λ values yield more smooth, slowly changing models. For fixed λ, the best-fit parameters can be found by least-squares. We identify the optimal value of λ for each light curve using generalized cross-validation (Golub et al. 1979; Craven & Wahba 1979). This allows the data for each light curve to drive the complexity of the model while also constraining the model to not overfit the data. Typical values of λ lie between 0.1 and 50 for our ACVS light curves, with an average value of 5.

3.1.2. Novel Light-curve Features

In addition to the 32 periodic and 20 non-periodic features used in Richards et al. (2011) to parameterize variable stars, we add 16 new features based on our generalized Lomb–Scargle periodogram, of which 10 were also used by Long et al. (2012). These features are compiled in Table 2. The first two features are `freq_amplitude_ratio_21` and `freq_amplitude_ratio_31`, which are ratios of the amplitudes of the second to first and third to first frequencies, respectively. The feature `freq1_lambda` is the optimized value of λ in Equation (3) found by generalized cross-validation. We also add three features aimed at detecting eclipsing sources from the Lomb–Scargle model in Equation (2), phased on twice the Lomb–Scargle period. We compute the phases and magnitudes of the two distinct minima and two distinct maxima of the phased light-curve model. The feature `freq_model_max_delta_mags` is the absolute value in the magnitude difference between the two model light-curve magnitude maxima (i.e., eclipses), and should be non-zero if the source is an eclipsing binary. Similarly, the feature `freq_model_min_delta_mags` captures the absolute value in the magnitude difference between the two magnitude minima and the feature `freq_model_phi1_phi2`, which is constructed to detect eccentric binary systems, is the ratio of the phase difference between the first minimum and the first maximum (i.e., primary eclipse) to the phase difference between the first minimum and second maximum (i.e., secondary eclipse).

Table 2
Light-curve Features Used in Addition to the Features of Richards et al. (2011)

Feature	Description
freq_amplitude_ratio_21	Amplitude ratio of the second to first Fourier component in the Lomb–Scargle model
freq_amplitude_ratio_31	Amplitude ratio of the third to first Fourier component in the Lomb–Scargle model
freq_model_max_delta_mags	Absolute value of mag difference between the two model light-curve maxima phased on 2P[a]
freq_model_min_delta_mags	Absolute value of mag difference between the two model light-curve minima phased on 2P
freq_model_phi1_phi2	Ratio of the phase difference between the first minimum and the first maximum to the phase difference between the first minimum and second maximum
freq_n_alias	Number of top period estimates that are consistent with a 1 day period
freq_rrd	Boolean that is true only if freq_frequency_ratio_21 or freq_frequency_ratio_31 are consistent with 0.746
freq1_lambda	Optimal value of λ from Equation (3) found by generalized cross-validation
gskew	$(\mathrm{med}(m) - \mathrm{med}(m[0:p])) + (\mathrm{med}(m) - \mathrm{med}(m[p:1]))$, where we choose $p = 0.03$
scatter_res_raw	MAD of the Lomb–Scargle residuals divided by the MAD of the raw light-curve values
p2p_scatter_2praw	Sum of squared mag differences between pairs of successive observations in the light curve folded around 2P divided by that of the raw light curve
p2p_scatter_over_mad	Median of the absolute differences between successive observations normalized by the MAD
p2p_scatter_pfold_over_mad	Median of the absolute differences between successive mags in the folded light curve normalized by the MAD of the raw light curve
medperc90_2p_p	90th percentile of the absolute residual values around the 2P model divided by the same quantity for the residuals around the P model
fold2P_slope_10percentile	10th percentile of slopes between adjacent mags after the light curve is folded on 2P
fold2P_slope_90percentile	90th percentile of slopes between adjacent mags after the light curve is folded on 2P
p2p_ssqr_diff_over_var	The sum of squared mag differences in successive measurements divided by the variance

Note. [a] We use P to denote the Lomb–Scargle estimated period, and 2P to be double that period.

Additionally, we introduce the feature freq_n_alias, which counts the number of frequency estimates that are consistent with the parasite frequency of 1 cycle per day.[10] This feature supplements the freq_signif feature (computed using the method of Zechmeister & Kürster 2009) to determine whether a source is, in fact, periodic. We further add the class-specific feature freq_rrd, which indicates whether the ratio of the first to second frequency is consistent with 0.744, which is the frequency ratio enjoyed by Double-Mode RR Lyrae variable stars (Szczygieł & Fabrycky 2007).

Finally, we add the following five features which are adopted from Dubath et al. (2011). The feature scatter_res_raw computes the ratio of the median absolute deviation (MAD) of the residuals of the Lomb–Scargle model to the MAD of the raw light curve. The features p2p_scatter_2praw,p2p_scatter_over_mad, and p2p_scatter_pfold_over_mad are the sum of squared differences of the scatter about the light curve phased on the Lomb–Scargle period to that of either the phased or raw light-curve data. Similarly, the feature medperc90_2p_p is the 90th percentile of the absolute residual values around the model phased on twice the Lomb–Scargle period divided by the same quantity for the residuals around the model phased on the Lomb–Scargle period. Furthermore, we develop two new features, fold2P_slope_10percentile and fold2P_slope_90percentile, which are the 10th and 90th percentile slopes of the Lomb–Scargle model around twice the period, intended to capture the steepness of the ingress and egress of eclipse. We also add a new measure of skew, gskew, which is a robust measure of skew designed to detect objects which have abrupt decreases in brightness. Lastly, we add the feature p2p_ssqr_diff_over_var from Kim et al. (2011), which is the sum of squared magnitude differences in successive measurements divided by the variance.

3.1.3. Correcting Eclipsing Periods

Comparison of our estimated periods with those from the ACVS catalog reveals that our period estimates are often exactly half of the ACVS period for sources that are classified as eclipsing binaries by ACVS. Of the 5913 objects that are classified as eclipsing binaries in ACVS, our period estimate matches the ACVS period for only 1353 sources (23%) and was exactly one-half of the ACVS period for 4184 sources (71%). After visual inspection of some of these light curves, we find that for eclipsing binaries in which our periods differ, the ACVS period is correct for most (but not all) of the objects. Using a visually confirmed set of 150 eclipsing sources in which our period is exactly one-half of the true period and 150 eclipsing binaries for which our period is correct, we construct a supervised machine-learned RF classifier on all of the features described in Section 3.1 to automatically discover, for each eclipsing source in the data set, whether our estimated period is correct or wrong by a factor of one-half.

In this classifier, the most important features (see Section 3.1.4) in determining whether our period is correct are, unsurprisingly, the freq_model_max_delta_mags, freq_model_min_delta_mags, and freq_model_phi1_phi2 features, which capture differences between the primary and secondary eclipses, and freq1_harmonics_amplitude_1, the amplitude of the first harmonic, which will be large for an eclipsing binary containing two unequal eclipse depths that was incorrectly identified as having period one-half of the true eclipsing period. We apply this classifier to all 11,138 sources in our data set that were either classified by ACVS as an eclipsing binary or whose most probable class from our variable star classifier was one of the eclipsing binary classes. Of those sources, the classifier determined that our period was correct for 5456

[10] Originally, we experimented with modifying the frequency estimates for objects whose principal frequency estimate was consistent with a parasite frequency. However, we found that this procedure was slightly detrimental to the classifier and that no significant artifacts occur due to the prevalence of objects at $f = 1, 2, 3$, etc., cycles per day.

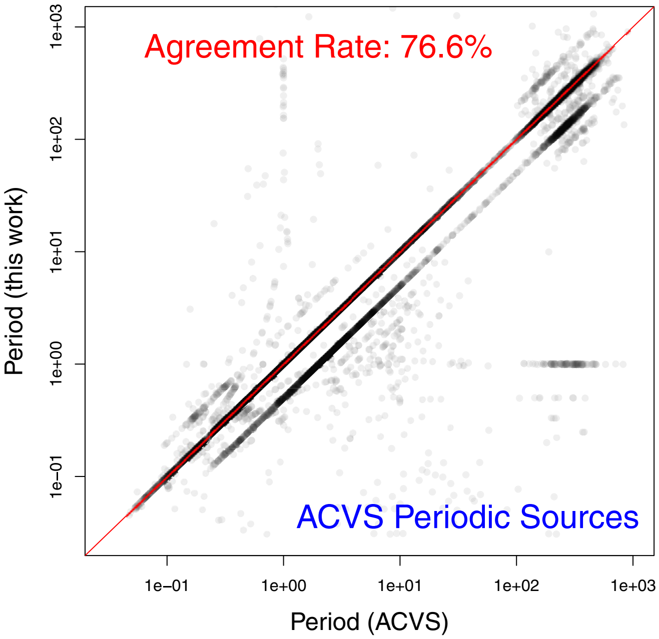


Figure 2. Period estimated by our period-finding algorithm vs. the period stated in the ACVS catalog, for all 12,008 ASAS periodic sources in ACVS. The red dashed line denotes perfect agreement; for a total of 9280 of the stars (76.6%) we find periods that exactly match the ACVS period. For 93.0% of these sources, our period estimate either matches the ACVS period exactly or is different by a factor of two.

(A color version of this figure is available in the online journal.)

objects and that our period was wrong by a factor of 1/2 for 5753 sources. Doubling the period of those 5753 sources yielded a significant boost in the period agreement rate with the ACVS eclipsing binary stars, with 4146 of 5913 (70%) of those sources resulting in a period match.

In Figure 2 we plot, for the 12,008 ASAS sources which the ACVS confidently classified into a single periodic class (i.e., not classified as "MISC" and not listed in multiple classes), the ACVS period versus our estimated period. Our agreement rate with ACVS is 76.6% on these objects. Including matches to half and twice the ACVS period yields an agreement rate of 93.0%. To evaluate the overall accuracy of our period finder, we chose a random 40 objects from the 7.0% of sources for which our period estimate differs from that of ACVS. Of these 40 sources, 23 were cases in which our period was obviously correct while only 3 were cases in which the ACVS period was obviously correct. The other 14 objects were inconclusive due to too few data or likely aperiodic or quasi-periodic variability (and should not have been identified as periodic by ACVS). All three sources for which our period is incorrect are detached eclipsing systems with very sharp eclipses. Hence, we are confident in the accuracy of our periods in all cases except detached eclipsing systems or other variables with sharp periodic features.

3.1.4. Feature Importance

We plot the RF importance measure of the top 20 features in the classification RF in Figure 3. The RF feature importance measure describes the decrease in overall classification accuracy that would result if the feature were replaced by a random permutation of its values. See Breiman (2001) for further details. In Figure 3 we find that the fundamental frequency of oscillation (i.e., period) of the light curve is by far the most important feature in the classifier. Other important features include estimates of the light-curve skew (captured by both `skew` and `gskew`), the degree of periodicity of the source (`scatter_res_raw`), the amplitude of the fundamental frequency, measurements of amplitude/variability (`stetson_j`,`std`,`median_absolute_deviation`), various colors, and features extracted from the light curve folded on twice the period. One caveat to the feature importance measure is that it does not account for correlations between features. For instance, the standard deviation and MAD of the light curve both provide measurements of the spread in the flux measurements about the average value; thus, the conditional importance of `std` given `median_absolute_deviation` is quite low even though their individual importance measures are both large. Dubath et al. (2011) account for this by iteratively removing features that are highly correlated with the most important features.

To determine the optimal set of features to include in the RF model, we run a pair of feature selection routines. First, starting with an empty feature set, we iteratively add the feature with the next-highest RF feature importance score. For each feature set, we compute the fivefold cross-validation misclassification rate over the 810 ASAS training objects (we use the combined training set to train the RF, but evaluate performance only on the ASAS sources). Results of this procedure are plotted in the left panel of Figure 4. We find that performance improves during the addition of the first ∼20 features and then levels off as the subsequent (less important) features are added to the model. Crucially, the performance never significantly deteriorates as more features are added. Second, we use the feature-selection routine of Dubath et al. (2011), where features are incrementally added to the model via their RF importance score, but only if they are not highly correlated ($<$0.8 Spearman correlation) with any feature already in the set. Results of this experiment, in the right-hand panel of Figure 4, are consistent with those of the other procedure, with the performance of the all-features model being consistent with the performance of the best overall model. Since expected error rates never significantly increase with additional features, we choose to use *all* 72 features in our model. Though the extra features are not informative for error rates, they may be useful for anomaly detection (Section 3.5) or may have an effect when measured by other performance metrics (e.g., area under the ROC curve).

3.2. Training the Classifier

Non-parametric supervised classification methods, such as RF, require a training set of data with known class label to learn the mapping from feature space to classes. Once this model is learned, data from each ASAS source can be trivially fed into the model to attain probabilistic classifications for each object. However, much care must be taken to attain a training set that is representative of the ASAS data. If significant discrepancies exist between the distribution of training features and the distribution of the features of the ASAS data, then, as shown by Richards et al. (2012), significant biases can occur in the ASAS classifications due to poor model selection and catastrophic errors caused by sample-selection bias. In this section, we detail the construction of our classification training set and efforts to avoid sample-selection bias.

As the base training set for the ASAS classifier, we use the training set of confirmed *Hipparcos* and OGLE sources used in Richards et al. (2011) (which is based on, but slightly different than, the training set used by Debosscher et al. 2007). This data set consists of 1549 variable stars from 27 different science

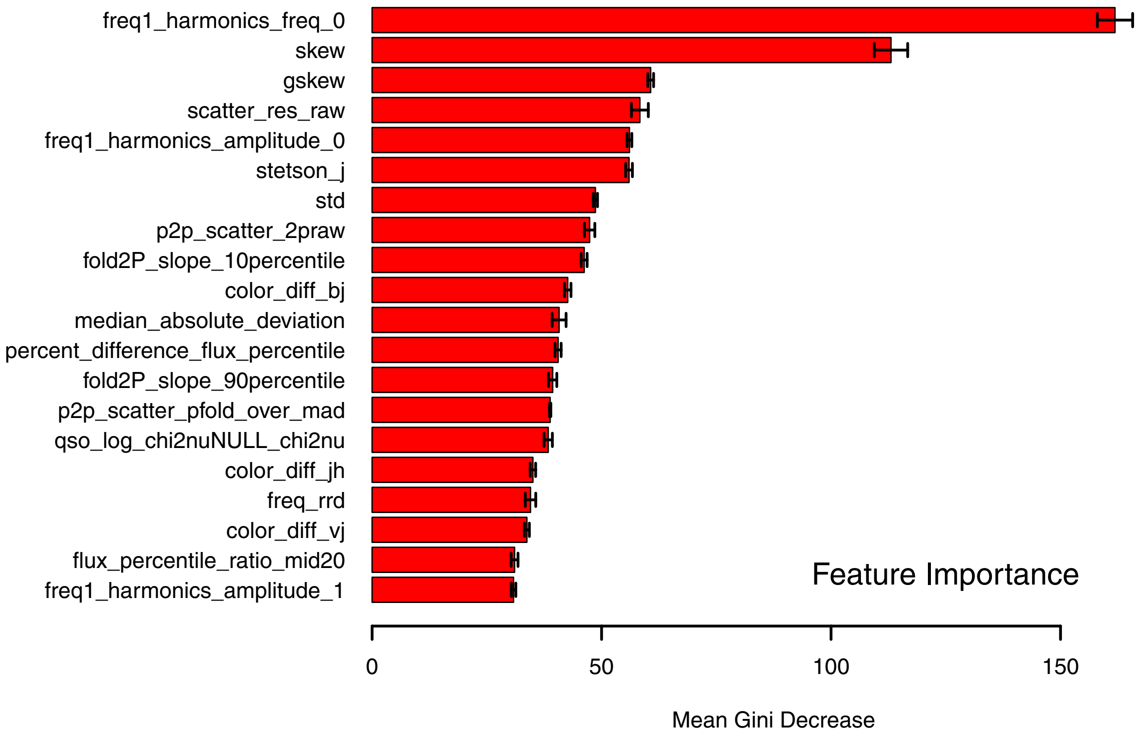


Figure 3. Random forest feature importance for the top 20 features, as estimated by calculating the mean feature importance over five random forest classifiers. As expected, the fundamental frequency of oscillation is the most important feature in ASAS variable star classification. The next most important features include the skew of the flux measurements, the ratio of the standard deviation of the scatter about the Fourier model to the raw observed scatter, the Fourier model amplitude of the fundamental frequency, and the Stetson variability index *J* (Stetson 1996). Error bars denote the standard deviation in the feature importance over five random forests (each initialized with a different random seed).

(A color version of this figure is available in the online journal.)

classes.[11] Next, we cross-match the *Hipparcos* training set with our ASAS sample, finding 266 matching sources. For these 266 training objects, we replace their *Hipparcos* light curves with their ASAS light curves in the training set. At this stage of the analysis, we also choose to exclude four variable star classes: Lambda Böotis, Slowly Pulsating B, Gamma Doradus, and Wolf-Rayet. Each of these classes of variable star is populated by objects whose amplitude of variability is $\Delta V \lesssim 0.05$ mag, which is below the ACVS variability selection cut of 95th percentile in the magnitude-dispersion diagram (Pojmański 2002). Indeed, of the 113 variable stars in our *Hipparcos* training set that belong to one of these four classes, not a single star passed the variability cuts used to construct the ACVS catalog, even though 78 of the 113 stars were observed by ASAS. Because such prototypical examples of each of the four small-amplitude classes did not satisfy the cuts used to construct ACVS, we do not expect to find any objects of these classes in the ACVS sample.

The feature distribution of this initial training set is substantially different than the bulk distribution of ASAS features (see Figure 1 of Richards et al. 2012). In Richards et al. (2012) it was shown that this mismatch causes poor performance by supervised machine-learned classification and demonstrated that an active learning framework could be used to supplement the training set in a statistically rigorous manner. Active learning is a classification paradigm in which the supervised classifier is able to query the human user for the classification labels of a subset of sources with unknown class, whereby these objects are manually labeled by the user and added to the training set. Using an RF classifier, the active learning query function S_2 from Equation (5) of Richards et al. (2012), and the crowd sourcing methodology outlined in that work, we add 407 ASAS sources to the training set.

In addition to the 407 active-learning training sources, we supplemented the classification training set with matched sources from the SIMBAD catalog (Wenger et al. 2000) using a combination of algorithmic catalog matching, literature searching, and human vetting. Starting with the list of NOMAD sources associated with ASAS sources (see Section 2.3), our algorithm looks for a SIMBAD source which is spatially close to the NOMAD source, calling a match any SIMBAD source which is within $0\farcs5$ of the NOMAD source. If no SIMBAD source fits this constraint, then no association is made. Our primary purpose for this exercise was to strengthen the training set for underrepresented science classes. Thus, for any positive SIMBAD association of class RV Tauri, Population II Cepheid, Beta Cephei, Chemically Peculiar, T Tauri, or Herbig Ae/Be, we performed a literature search on the object, only including the source in the training set if it was definitely confirmed by multiple sources. This procedure allowed us to add 68 sources to the training set. At this point, we also added R Coronae Borealis (RCB)—a class of hydrogen-deficient carbon-rich supergiants that undergo episodes of extreme dimming (Clayton 1996)—to the training set, populating the training sample with 17 RCB stars found via the SIMBAD matching procedure.

In a preliminary edition of the classification catalog it was noticed that an excessively large fraction of the ACVS variables, $\gtrsim$10%, were being classified as T Tauri stars (TTS). At the time TTS only constituted $\sim$0.7% of the training set so the

[11] Note that this training set is slightly different than that of Richards et al. (2011) in that we further split the T Tauri class into Classical (nine stars) and weak-line (two stars) subclasses and add the SX Phoenicis variable class.

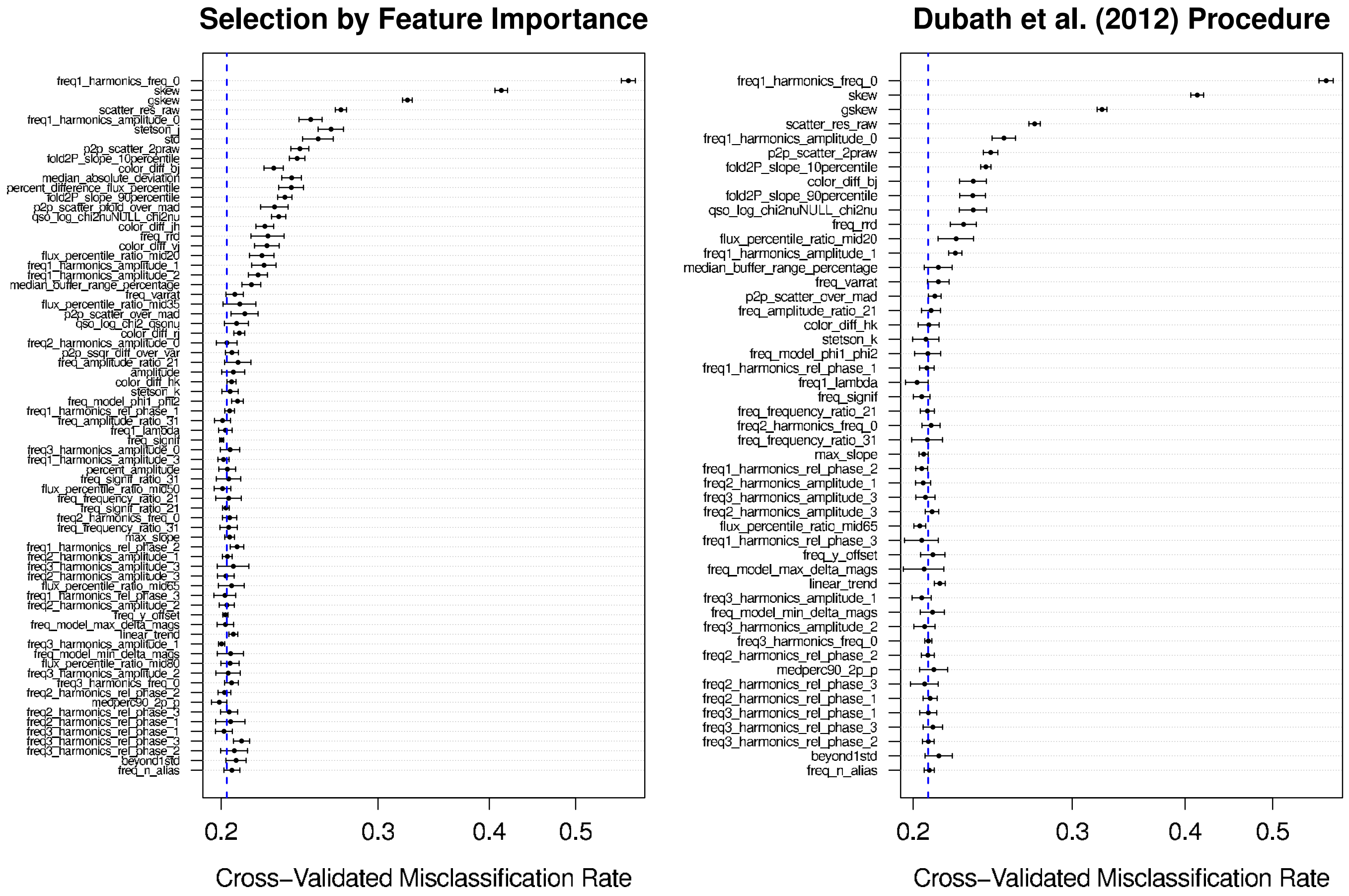


Figure 4. Result of feature selection experiments. Left: starting with an empty feature set, features are added in order of their random forest feature importance score. As more features are added, the cross-validated error rate over ASAS training data decreases and eventually levels off. The vertical blue dashed line marks 1σ above the lowest average cross-validated error rate of any single feature set. Right: in a similar experiment, features are again added in order of RF importance, but only if they are not highly correlated with a feature already in the set of used attributes. Qualitatively, the results of the two experiments are similar, and give us confidence that all features can be included in the model without overfitting.

(A color version of this figure is available in the online journal.)

large fraction of TTS classifications was not expected. Upon further inspection we discovered that the inclusion of the two subclasses of TTS, which exhibit significantly different photometric behavior, into a single class led to their significant overrepresentation in the final catalog. This occurred because the single super-class of TTS was not tightly clustered in feature space, with training examples ranging from high-amplitude types with variability due to active accretion to quiescent types exhibiting low-amplitude variability due to rotation of spots. This heterogeneity effectively allowed the RF classifier to allocate a large partition of feature space to the TTS class, which resulted in high TTS probability being assigned to a large number of stars.

Thus, we decided to split the TTS class into two classes: weak-line T Tauri stars (WTTS) and classical T Tauri stars (CTTS). This split is physically motivated as WTTS are older young stellar objects whose photometric variability is periodic and characterized by the rotational modulation of cool spots on the stellar surface; CTTS, on the other hand, are younger stars that are still actively accreting from a disk with a variability signature that is typically more chaotic than WTTS (for a review of TTS variability see Herbst et al. 1994 and references therein). To populate these two new classes we divided all members of the original TTS training set as well as new TTS identified via our SIMBAD–ASAS matching query, which included SIMBAD matches of type Y*O, Or*, pr*, or TT*.[12] We split these sources into the CTTS and WTTS classes using the classical diving line between the two: for CTTS the equivalent width (EW) of Hα emission is >10 Å, while for WTTS $\mathrm{EW}_{\mathrm{H}\alpha}$ <10 Å (see, e.g., Walter 1986; Strom et al. 1989). Stars were only included in the training set if we could find a published value of $\mathrm{EW}_{\mathrm{H}\alpha}$, which typically came from the catalogs of Herbig & Bell (1988) or Torres et al. (2006).

It was later noticed that several known members of the RS Canum Venaticorum (RS CVn) class of binary stars were being classified as WTTS, which prompted us to add RS CVn stars as a new class in the training set. To populate the RS Canum Venaticorum class in the training set we identified matches between ACVS sources and the catalog of chromospherically active binary stars (CABS; Strassmeier et al. 1988). The CABS includes both RS CVn and BY Draconis (BY Dra) binaries, both of which we include in the training set as the latter is the low mass analog of the former. In practice RS CVn and BY Dra stars exhibit the same photometric behavior, from a classification standpoint they can only be separated spectroscopically which is why we include them as a single class in the MACC. The cross-match between the CABS and ACVS produces 16 RS

[12] Y*O: Young Stellar Object; Or*: Variable Star of Orion Type; pr*: Pre-main sequence Star; TT*: T Tau-type Star.

Table 3
Class Distribution of Training Set Objects Used to Fit the Probabilistic ASAS Classifier

Science Class	N_{Train}	Prior $\mathbf{P}$(Class)
Mira	164	0.0852
Semireg PV	101	0.0525
SARG A	15	0.0078
SARG B	29	0.0151
LSP	54	0.0281
RV Tauri	25	0.013
Classical Cepheid	204	0.106
PopII Cepheid	27	0.014
Multi-Mode Cepheid	98	0.0509
RR Lyrae FM	148	0.0769
RR Lyrae FO	39	0.0203
RR Lyrae DM	59	0.0306
Delta Scuti	133	0.0691
SX Phe	6	0.0031
Beta Cephei	55	0.0286
Pulsating Be	49	0.0255
RSG	35	0.0182
ChemPeculiar	75	0.039
RCB	17	0.0088
ClassT Tauri	12	0.0062
Weak-line T Tauri	20	0.0104
RS CVn	17	0.0088
Herbig AEBE	22	0.0114
S Doradus	7	0.0036
Ellipsoidal	13	0.0068
Beta Persei	178	0.0925
Beta Lyrae	202	0.1049
W Ursae Maj	121	0.0629

Note. This class distribution defines the prior on class probabilities used to compute posterior class probabilities for each source.

CVn and 1 BY Dra, which we use to define the RS CVn training set.

Finally, we choose to replace the *Hipparcos* class of periodically variable supergiants with the more well-defined class of red super giants (RSG). RSGs in the Large and Small Magellanic Clouds are bright enough to be detected by ASAS, and they typically exhibit large amplitude ($\gtrsim$ few $\times$ 0.1 mag) variability leading to their inclusion in the ACVS. The class was identified as one with a substantial population during the search for new R Coronae Borealis stars (see Miller et al. 2012). The training set for the class consists of 35 stars which are spectroscopically identified as low gravity and have either a K or M spectral type as identified in Massey & Olsen (2003), Neugent et al. (2012), and references therein. A few additional RSGs, which we require to match the same spectroscopic criteria, were included following their identification during the search presented in Miller et al.

Our final training set consists of 1925 sources in 28 science classes. A total of 810 of these sources are observed by ASAS, so we use their ASAS light curves to derive features that we will use to train the classifier. For the other 1115 training objects, we only have data in *Hipparcos* (591 stars) or OGLE (524 stars), so we employ the light curves observed by those missions. A tabulation of the entire training set, by class, is given in Table 3. The implicit class prior in fitting an RF classifier is the empirical vector of training-set class proportions, which is given in Table 3.

Finally, we find the optimal RF model by minimizing the tenfold cross-validation classification error rate over the ASAS training set with respect to the number of RF trees, `ntree`, the number of features considered on each splitting node, `mtry`, and the minimum size of each terminal node, `nodesize`. Performing a grid search over those three parameters, we find that the optimal model is `ntree` = 10000, `mtry` = 23, and `nodesize` = 5, attaining an average 10-fold cross-validation error rate of 19.75% for the 810 ASAS training objects. The cross-validated confusion matrix for only the 810 ASAS sources is plotted in Figure 5. Here, we find at least 90% accuracy for Mira, Classical Cepheid, RR Lyrae, FM, RR Lyrae, FO, and Chem. Peculiar subclasses and at least 70% correspondence for 14 of the 28 classes. The classes for which we find less correspondence are those that have fewer than 10 ASAS training sources or are easily confused with other classes (e.g., SARG A versus B). For the remainder of this paper, and to construct the ASAS classification catalog, we use an RF trained on all 1925 training set objects with the optimized tuning parameters.

3.3. Calibrating Classifier Probabilities

Using the features described in Section 3.1 and the training set outlined in Section 3.2, we fit an RF classifier with optimized tuning parameters and use it to generate class predictions and full 28-class probability vectors for all 50,124 ASAS objects. A desirable property of probabilistic classifications is that they be *calibrated*. That is to say, if we consider all sources whose class probabilities for a particular class are 90%, then 90% of those objects should truly be of that class. Calibration is attractive because it allows us to treat the probabilistic classifier output as if it were truly a set of posterior class probabilities, $\mathbf{P}(\mathrm{class} \mid \mathbf{x})$. Calibration also allows us to easily substitute different prior class probabilities by multiplying the classification probabilities by the appropriate vector of prior ratios and re-normalizing the probability vectors (see Section 4.1 for a detailed explanation).

However, the class probabilities estimated by the RF are not necessarily calibrated. To check their calibration we perform the following experiment. Using only the subset of ASAS training data (810 objects), we perform tenfold cross-validation to estimate the RF classification probabilities for each source.[13] This provides a vector of 28 cross-validated class probabilities for each object. Then, in each of eight disjoint probability bins (chosen such that each bin contains at least 100 instances), we compute the proportion of the objects, p_{true}, that are truly of the specified class. If the probabilities were calibrated, then the value of p_{true} should match the mean RF probability within each bin. In Figure 6 we see, by the solid black line, that this certainly is not the case for our classifier. Specifically, the RF classifier tends to be conservative in that it systematically estimates a smaller probability than p_{true} for the RF probabilities greater than $\sim$0.3. For instance, in the RF probability bin centered around 0.5, around 70% of those objects are truly of the specified class.

Two popular methods exist for calibrating classifier probabilities using simple transformations. Platt Scaling (Platt 1999) transforms the probabilities using a sigmoid function whose parameters are chosen via maximum likelihood over the training set. Isotonic Regression (Robertson et al. 1988; Zadrozny & Elkan 2001) is more flexible, replacing the sigmoid function

[13] Cross-validation ensures that each object is held out of the training set when fitting the classifier that is used to predict the class probabilities for that object. In this sense, the cross-validated classification probabilities are representative of the classifier probabilities for the unlabeled data.

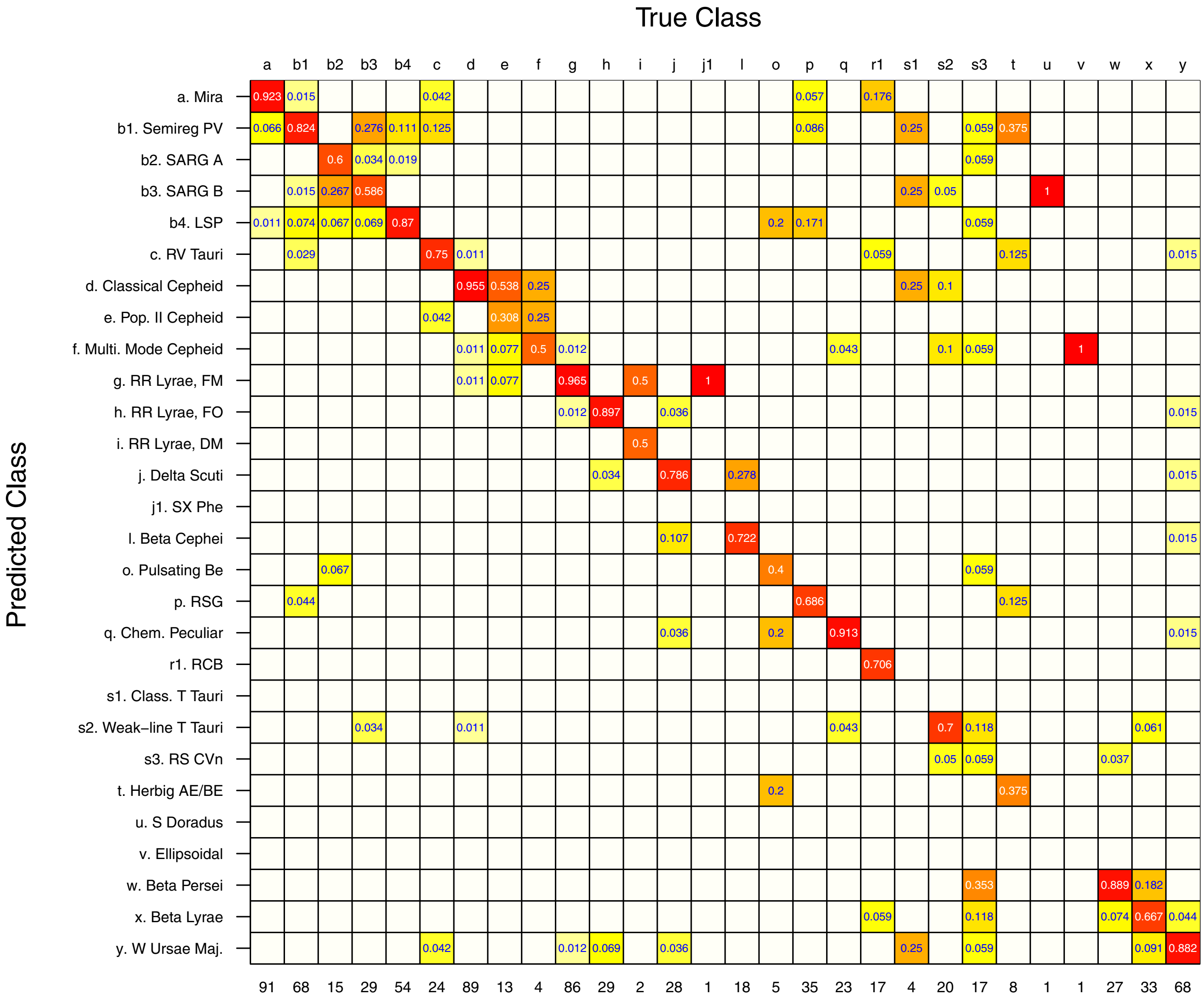


Figure 5. Cross-validated confusion matrix for all 810 ASAS training sources. Columns are normalized to sum to unity, with the total number of true objects of each class listed along the bottom axis. The overall correspondence rate for these sources is 80.25%, with at least 70% correspondence for half of the classes. Classes with low correspondence are those with fewer than 10 training sources or classes which are easily confused. Red giant classes tend to be confused with other red giant classes and eclipsing classes with other eclipsing classes. There is substantial power in the top-right quadrant, where rotational and eruptive classes are misclassified as red giants; these errors are likely due to small training set size for those classes and difficulty to classify those non-periodic sources.

(A color version of this figure is available in the online journal.)

with any monotonically increasing function (which is typically restricted to a set of non-parametric isotonic functions, such as step-wise constants). A drawback to both of these methods is that they assume a two-class problem; a straightforward way around this is to treat the multi-class problem as C one-versus-all classification problems, where C is the number of classes. However, we find that Platt Scaling is too restrictive of a transformation to reasonably calibrate our data and determine that we do not have enough training data in each class to use Isotonic Regression with any degree of confidence.

Ultimately, we find that a calibration method similar to the one introduced by Bostrom (2008) is the most effective for our data. This method uses the probability transformation

$$\widehat{p}_{ij} = \begin{cases} p_{ij} + r(1 - p_{ij}) & \text{if } p_{ij} = \max\{p_{i1}, p_{i2}, \ldots, p_{iC}\} \\ p_{ij}(1 - r) & \text{otherwise,} \end{cases} \tag{4}$$

where $\{p_{i1}, p_{i2}, \ldots, p_{iC}\}$ is the vector of class probabilities for object i and $r \in [0, 1]$ is a scalar. Note that the adjusted probabilities, $\{\widehat{p}_{i1}, \widehat{p}_{i2}, \ldots, \widehat{p}_{iC}\}$, are proper probabilities in that they are each between 0 and 1 and sum to unity for each object. The optimal value of r is found by minimizing the Brier score (Brier 1950) between the calibrated (cross-validated) and true probabilities.[14] We find that using a fixed value for r is too restrictive and, for objects with small maximal RF probability, it enforces too wide of a margin between the first- and second-largest probabilities. Instead, we implement a procedure similar to that of Bostrom (2008) and parameterize r with a sigmoid function based on the classifier margin, $\Delta_i = p_{i,\max} - p_{i,\text{2nd}}$, for each source,

$$r(\Delta_i) = \frac{1}{1 + e^{A\Delta_i + B}} - \frac{1}{1 + e^{B}}, \tag{5}$$

where the second term ensures that there is zero calibration performed at $\Delta_i = 0$. This parameterization allows the amount of

[14] The Brier score is defined as $B(\widehat{p}) = 1/N \sum_{i=1}^{N} \sum_{j=1}^{C} (I(y_i = j) - \widehat{p}_{ij})^2$, where N is the total number of objects, C is the number of classes, and $I(y_i = j)$ is 1 if and only if the true class of the source i is j.

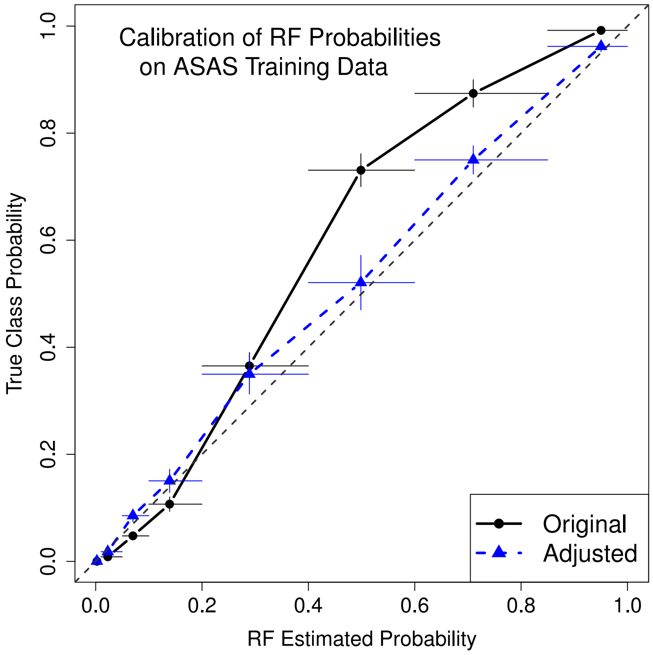


Figure 6. Reliability diagram for ASAS training data. The closer the curve follows the diagonal, the better calibrated the classifier probabilities. The initial random forest probabilities (solid black line) are not well calibrated, as the cross-validated ASAS RF probabilities tend to grossly underestimate the true posterior probabilities for large estimated probabilities. Using the calibration procedure of Bostrom (2008) results in well-calibrated adjusted probabilities (dashed blue line), as they are consistent with the diagonal of the reliability diagram. In the final MACC catalog, we use this calibration procedure to adjust all of the posterior probability estimates.

(A color version of this figure is available in the online journal.)

calibration adjustment to differ between objects with confident (high-margin) and ambiguous (low-margin) classifications. We choose the optimal value for the parameters A and B by minimizing the Brier score for the cross-validated classifications on the 810 ASAS training set data. Indeed, as expected, we find that the proper amount of adjustment is low for stars with small RF margin (e.g., $r(0.05) = 0.10$) and higher for sources with large RF probability margin (e.g., $r(0.5) = 0.57$). The parameters that minimize the Brier score over the training set are $A^* = -8.30$ and $B^* = 0.37$.

With the Bostrom (2008) calibration procedure, we correct the RF probability estimates for all ASAS sources. To test the efficacy of our procedure, we plot, in the blue dashed line in Figure 6, the adjusted (cross-validated) RF probabilities versus true posterior probabilities for our set of 810 ASAS training set objects. The calibration is now substantially improved over the raw RF probabilities as the calibrated probabilities are consistent with the true posterior class probabilities. Note that the adjusted probabilities are still slightly conservative in that, on average, the estimated probabilities are systematically smaller than the true probabilities for estimated probabilities greater than ∼0.1. In Figure 7, we plot these reliability diagrams for each of four subclasses of variable stars. Within each of the four subclasses, the calibration has improved, with marked decrease in the Brier score for each subclass; large deviation in one of the Eruptive + Rotational bins occurs due to low number statistics, with only four objects falling in that particular bin.

3.4. *Difficult Class Boundaries*

There are certain classes of variability that are difficult to separate based on photometric information alone. For instance, W Ursae Majoris, Delta Scuti, and RR Lyrae, FO stars all show variability on the same timescales with similar amplitudes. Other classes such as weak-line T Tauri and RS CVn stars exhibit variability from similar physical mechanisms (in this case, rotation of chromospherically active stars), which may result in ambiguous classification of sources of those classes based on light-curve information alone. An advantage of using machine-learned classification is that, given enough training data, these methods can learn which light-curve features best separate sources of similar class and can determine optimal class boundaries. In Figure 8, we plot the most informative features for separating notoriously difficult-to-separate classes of variable star. Even with relatively few training instances, the classifier effectively learns how to best distinguish, e.g., Delta Scuti stars and Beta Cephei stars.

That said, there will always be borderline cases, for which, given their light-curve and color data, it is impossible to confidently place the objects into a class. This uncertainty is reflected by low posterior class probabilities, typically $\lesssim 0.3$, assigned by the classifier across all classes. In Figure 9, we plot the ASAS light curves for a few of the least confidently classified (lowest maximal posterior probability) sources in MACC. These sources typically have poor data quality and/or fall in outlying regions of light-curve feature space, meaning that there is not enough light-curve information from these objects for the classifier to make a confident statement about their class. For comparison, in Figure 10 we plot a few of the ASAS objects whose light curves have a low anomaly score (see Section 3.5) but whose highest posterior class probability is smaller than 0.5. These light curves do not show atypical behavior, but tend to reside on the boundary between classes. Most of these objects reside on either the W UMa–Delta Scuti locus or between SARG A and B subtypes, making them impossible to classify with any degree of confidence. Likewise, ASAS 064635−1455.5 resides near the boundary between Delta Scuti and SX Phoenicis variability, which is difficult to disambiguate without metallicity measurements. Also, ASAS 210538+2005.0 is a Cepheid with atypically high amplitude and short period that places them near the dividing line between Classical and Population II Cepheid stars.

3.5. *Detecting Anomalous Objects*

Our calibrated ASAS probabilistic classification catalog supplies, for each object, its posterior probability of belonging to each of 28 science classes given its observed ASAS light curve and colors. These posterior class probabilities assume prior class probabilities given by the distribution of object types in the training set (see Table 3). The posterior probabilities also assume that the training set is fully representative of the set of ASAS data, meaning that all classes present in the ASAS data are represented in the training data and that the distribution of ASAS features is the same as the training set feature distribution. However, there is no guarantee that these conditions will be satisfied for each ASAS object, even after performing several rounds of active learning to reduce the discrepancies between the training and ASAS data sets.

The challenge, then, is to identify ASAS objects that do not resemble any of the training data. Classifier predictions for these objects will be dubious due to the outlying nature of their feature vectors compared to the training set feature distribution, either due to their belonging to a class not included in the training set or anomalous features brought about by noise or atypical physical variability. To detect such anomalies, we compute,

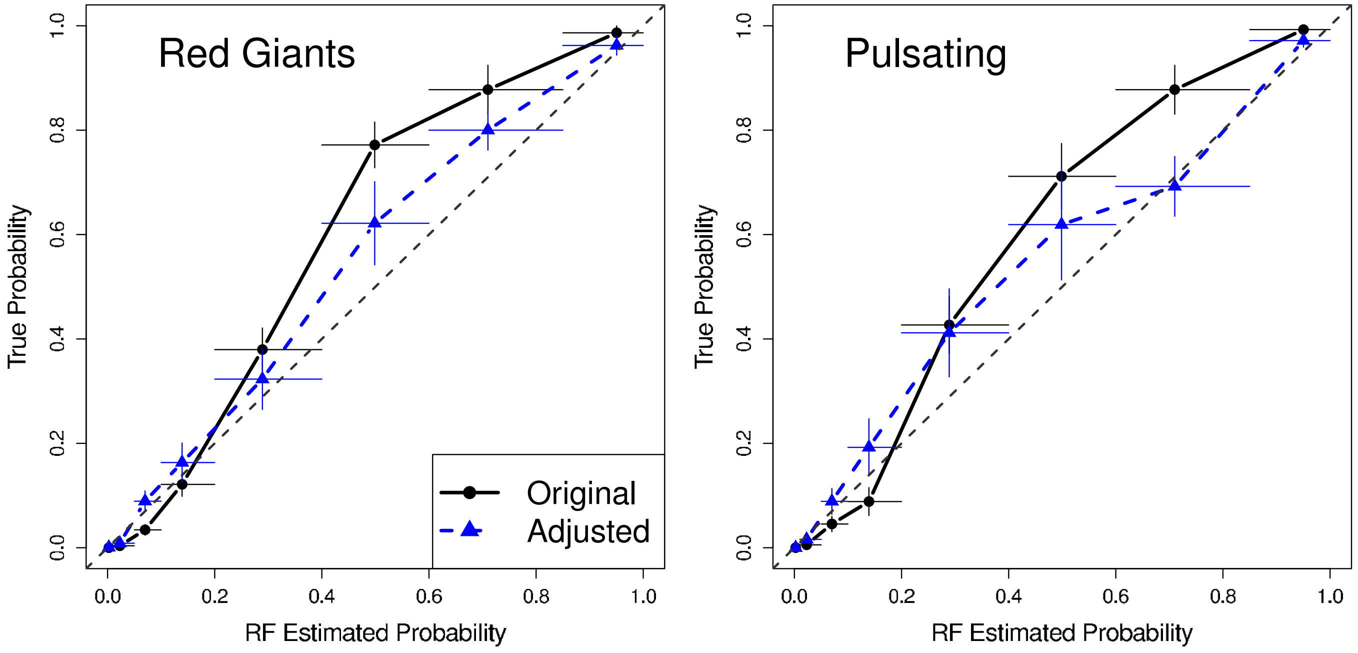
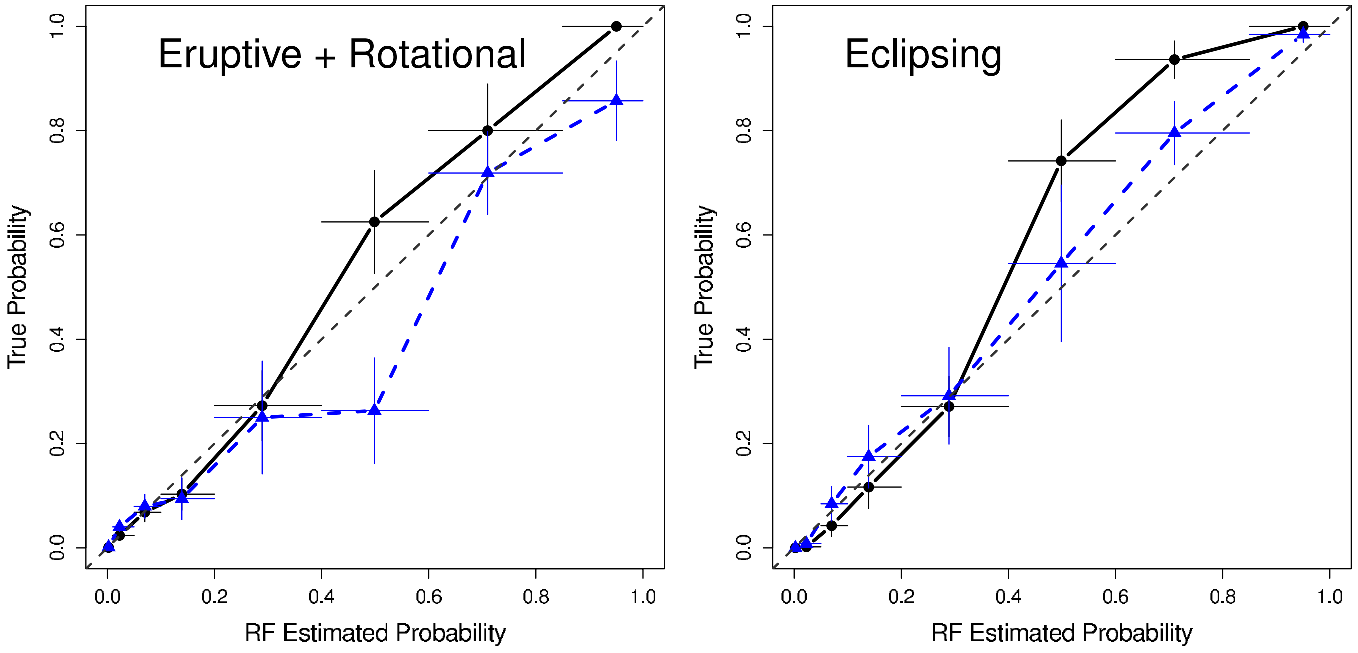



Figure 7. Reliability diagrams for each of the four subclasses in the ASAS training data. Within each subclass, the calibration procedure (dashed blue lines) produces better calibration than the raw, uncalibrated random forest estimates (solid black lines) in terms of Brier score. Whereas the off-the-shelf random forest probabilities are systematically too conservative for large estimated probabilities within each of the four subclasses, the adjusted probabilities are more consistent with the diagonal for almost every probability bin.

(A color version of this figure is available in the online journal.)

for each ASAS object, a distance metric from that object's feature vector to each source in the training set. In contrast to previous methods, which compute distances between phased light curves for periodic variable stars to detect anomalies (Protopapas et al. 2006; Rebbapragada et al. 2009), we compute a distance measure between feature vectors.

Similar to Bhattacharyya et al. (2011), we use a semi-supervised approach to compute the anomaly score for each variable star. We begin by fitting an RF classifier to the training set as in Section 3.2. The RF outputs a proximity measure ρ_{ij}, between each pair of sources i and j, which gives the proportion of trees in the RF for which the feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ appear in the same terminal node. If two sources have similar feature vectors with respect to the topology of the RF, then the proximity will be near 1, whereas if the feature vectors are dissimilar then the proximity will be near 0. Using the proximity measure, we define the discrepancy between the two feature vectors $\mathbf{x}_i$ and $\mathbf{x}_j$ as

$$d(\mathbf{x}_i, \mathbf{x}_j) = \frac{1 - \rho_{ij}}{\rho_{ij}}, \qquad (6)$$

which takes on non-negative real-valued numbers. This metric is semi-supervised because it uses the labeled training set

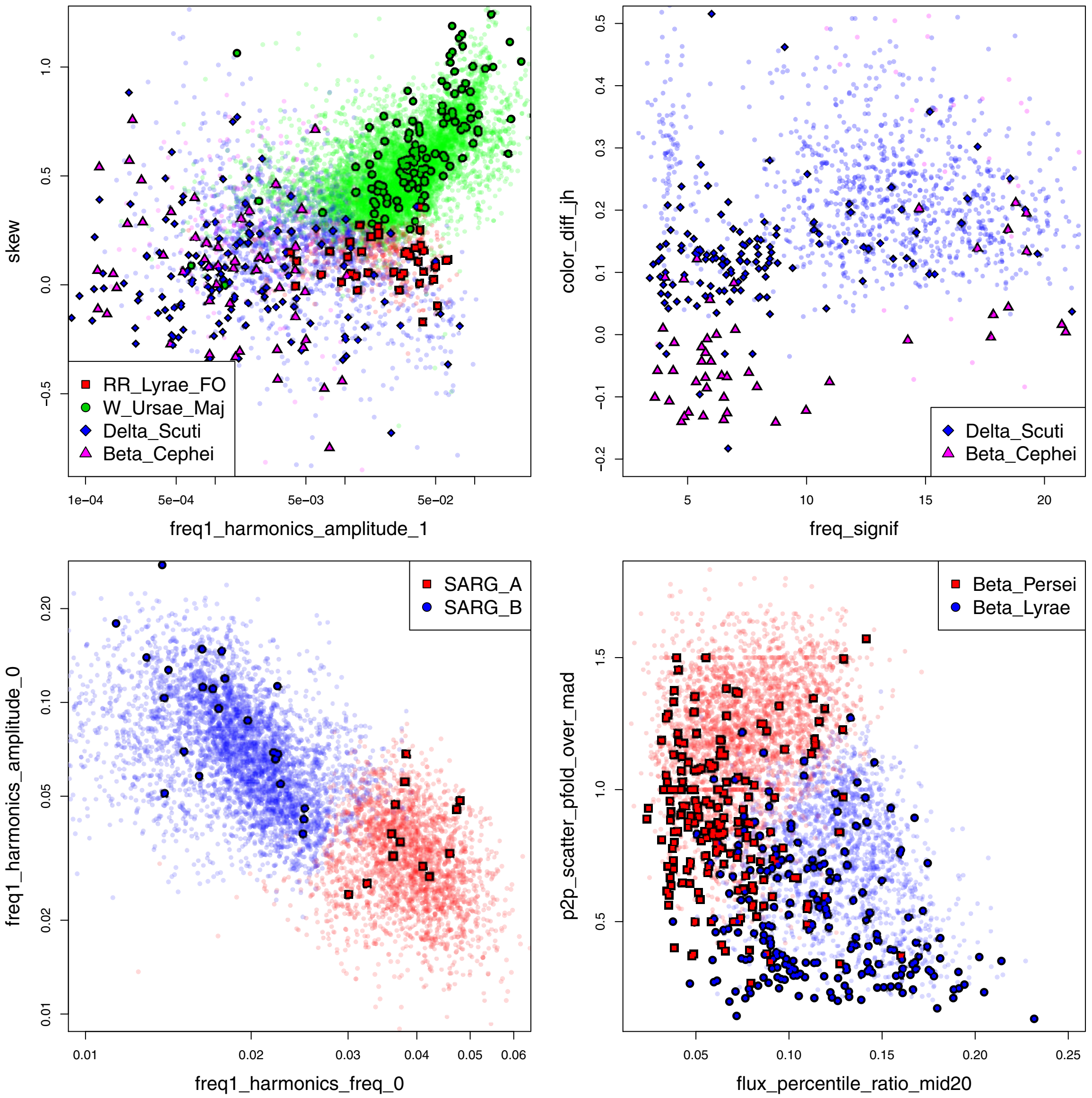


Figure 8. Random forest classifier automatically discovers class boundaries in the high-dimensional feature space. For certain easily confused classes, we plot the projections, in two-dimensional feature spaces, of training objects (points with solid outline) and MACC-classified objects (small dots). Top left: in the skew (first-harmonic-amplitude plane), W Ursae Majoris, RR Lyrae FO, and Delta Scuti stars are well separated, but Delta Scuti and Beta Cephei remain confused. Top right: however, Delta Scuti and Beta Cephei are separated by their $J - H$ color. Bottom left: SARG A and B subtypes split naturally in the period–amplitude plane. Bottom left: Beta Persei and Beta Lyrae binaries are largely separable by two features, with a small amount of overlap.

(A color version of this figure is available in the online journal.)

to construct the optimal RF classifier, which is then used to compute proximities (and discrepancies) between labeled and unlabeled sources.

The novelty of the distance measure in Equation (6) is that it automatically gives more weight to features that are important in the classifier while ignoring useless features. For instance, if a feature is important for classification, then the RF trees will make many splits on that feature, thus dividing the coordinate into many sub-regions. Hence, for a new source, the value of that class-predictive feature will have a great deal of power in determining which terminal node the source falls into for each tree, and thus will be a strong determinant of its proximity to other sources. Likewise, features that are unimportant for classification will never be split on by any tree, and thus proximities will be unaffected by their values. Unlike Euclidean distance, the proximity-based distance measure adapts to the geometry of the classification problem and can treat different regions of feature space differently based on the class boundaries and prevalence of training data in those regions.

Using the RF proximity measure, we construct an anomaly score for each ASAS object. We first compute the distance, using Equation (6), from the feature vector of each ASAS source to the feature vector of every training source. We define the anomaly score for each ASAS object to be the distance (Equation (6)) to the second nearest neighbor in the training set. Objects with large anomaly scores should be considered as outliers and their classifications should not be trusted because there is too much discrepancy between the features of those sources and the training set of variable stars. Note the subtle difference between the anomaly score and classification probability: sources with small maximal class probability may reside near training data but fall in regions of feature space

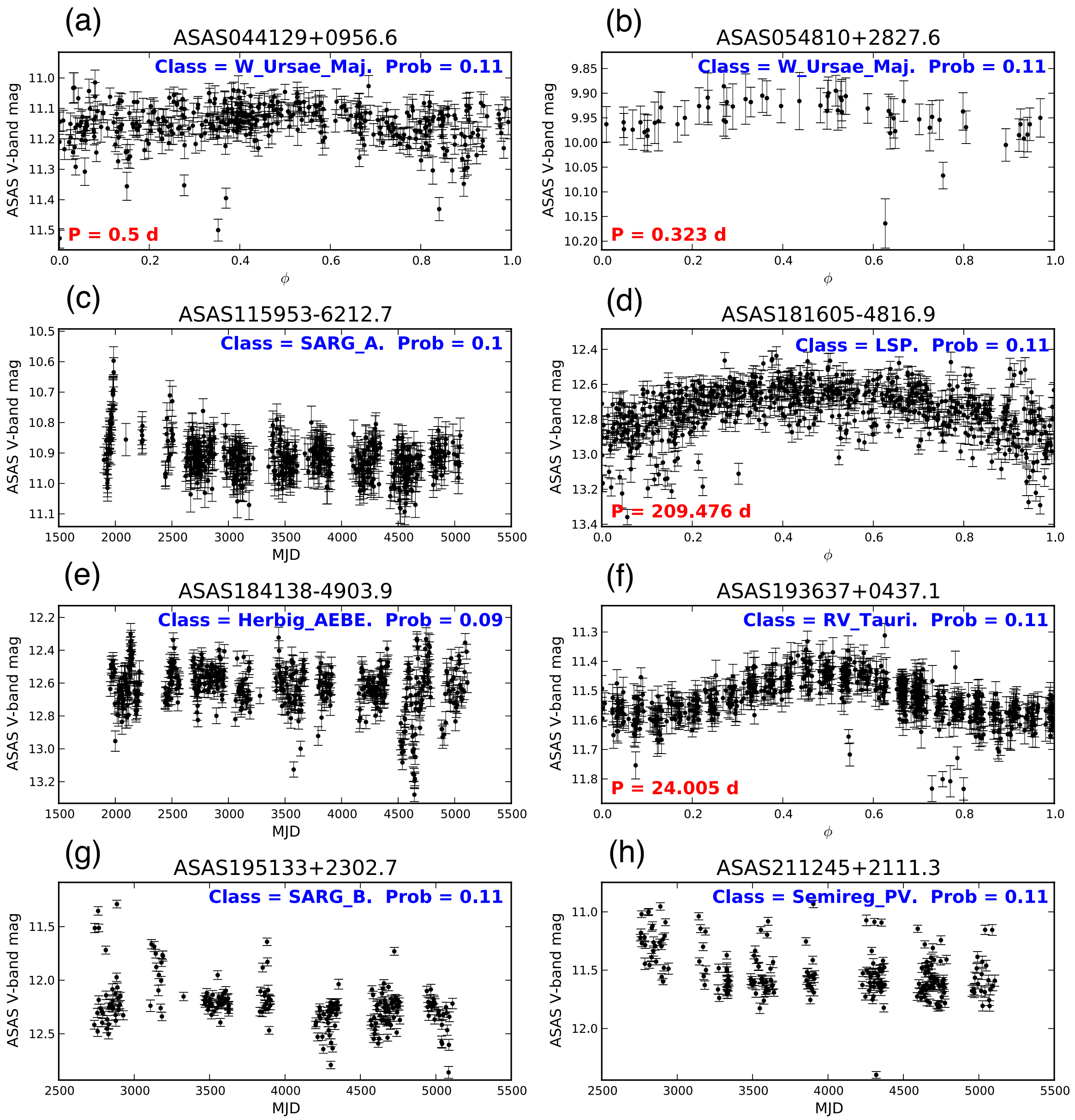


Figure 9. ASAS light curves for the candidates with lowest classification probability across all 28 classes. Several of these light curves suffer from lack of data (b), large temporal gaps (g), or large amounts of noise caused either by blending with nearby stars (a, c) or relative faintness (e). Others are outliers due to abnormal period–amplitude combinations (d, f), or secular variability on several year timescales (h). These objects, and others that obtain low probabilities across all 28 science classes, require further study to ascertain their true nature.

(A color version of this figure is available in the online journal.)

shared by several science classes. At the same time, sources with a high anomaly score may have a large maximal class probability due to their relatively close proximity to the training objects of a certain class compared to the training objects of the other classes.

The anomaly score provides a positive real-valued number for each object. However, we may ultimately want to make a decision, for each object, of whether or not that source is an outlier, by thresholding on the anomaly score. To determine an appropriate score threshold for anomaly detection, we employ cross-validation on the training set. In each of $K = 10$ cross-validation folds, we hold out a random subset of the ASAS training data, fit the RF classifier on the remaining data, and compute the anomaly score for each held-out object. Then, for each anomaly score threshold, we record the cross-validated classification error rate over the ASAS training data, counting each object whose anomaly score surpasses the threshold as an error. Results of this experiment are in Figure 11. As the threshold decreases, we identify more objects as outliers, but the classification error rate only becomes significantly affected for thresholds smaller than 10.5. Following the 1σ rule of Hastie et al. (2009) over 10 repetitions of the procedure, we find that the optimal threshold level is $t^* = 10.5$. Therefore, we recommend that the 1271 ASAS objects with anomaly scores larger than 10.5 be treated as outliers.

In Figure 12 we plot the ASAS light curves of eight sources that are amongst the highest outlier scores. These objects include a light curve showing rare year-long periodicity with small amplitude (ASAS103706−6528.3), a known pulsating Be star showing a high level of activity (ASAS143429−6412.1),

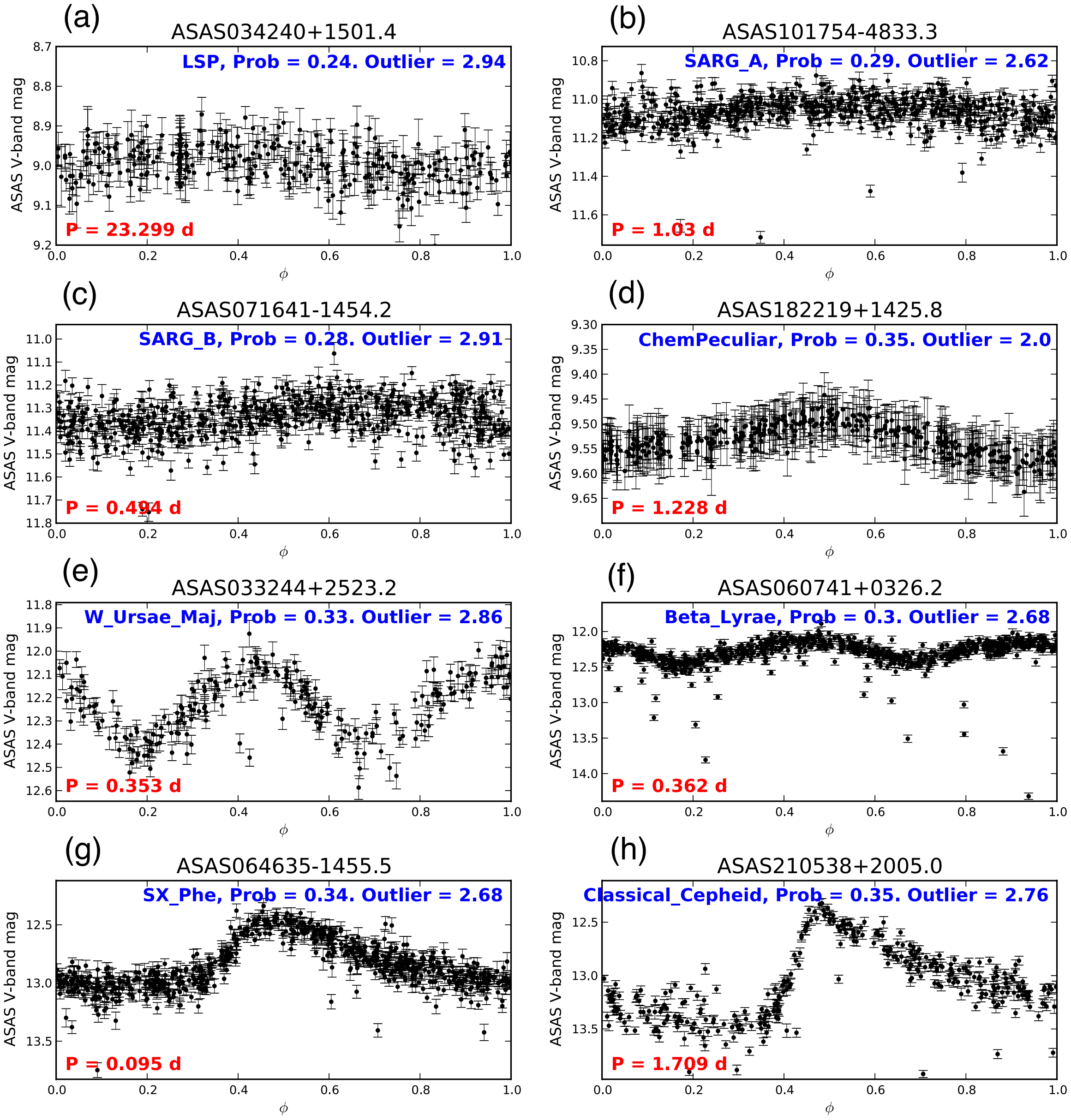


Figure 10. Light curves of ASAS objects whose anomaly scores are small even though their maximal classification probabilities are smaller than 0.5. These light curves show behavior that is not inconsistent with a particular class of variability, but typically reside between classes. The objects in (a, b, c) reside on the border between SARG A and B subtypes. The star in (d) has almost equal probability of being a Chemically Peculiar and W Ursae Majoris eclipsing variable. Likewise, the stars in (e–f) could either be eclipsing variables or Delta Scuti pulsating stars. The object in (g) lies on the boundary between SX Phoenicis and Delta Scuti, while that in (h) lies on the short-period end of the Cepheid locus, and is likely a Population II Cepheid.

(A color version of this figure is available in the online journal.)

another likely Be star showing semi-regular pulsations with amplitude modulation (ASAS073246−1519.3), and a star with very small amplitude 18.44 day periodicity (ASAS185203−2937.7). The other outliers in Figure 12 all have aperiodic variability with high-amplitude outbursts up to 1.5 mag in amplitude. For each of these outliers, there are no training instances that capture the observed variability in their ASAS light curves.

4. THE CATALOG

Here we describe the contents of the publicly available MACC. MACC is available for download at www.bigmacc.info and is also available in the online version of this publication. The first 30 rows of the classification catalog are reproduced here in Table 4. The columns of the catalog are as follows:

1. `ASAS_ID`: ID from ACVS
2. `dotAstro_ID`: ID from the online database http://dotastro.org/
3. `RA,DEC`: coordinates from ASAS[15]
4. `Class`: most probable class from the machine-learned classifier
5. `P_Class`: posterior probability that the source is from that class, given the ASAS light curve and colors
6. `Anomaly`: metric from Section 3.5; objects with a score greater than 10.0 should be considered as outliers
7. `ACVS_Class`: classification from the ACVS (Pojmański 2002)

[15] Coordinates from ASAS are sometimes wrong by several arcseconds due to its ∼15 arcsec pixel size. This effect is worse in crowded fields.

Table 4
The Machine-learned ASAS Classification Catalog

ASAS ID	dotastro[a]	R.A.[b]	Decl.[a]	Class	P(Class)	Anomaly	ACVS Class	Train Class	P(Mira)	. . .	P	P signif[c]	N	V[d]	ΔV[e]
000006+2553.2	215153	0.027375	25.886453	Mira	1	0.07	MIRA	Mira	1	. . .	319.295	10.398	170	8.56	2.488
000007+1844.3	215154	0.030375	18.738077	Beta Persei	0.788	2.012	ESD/CW-FU/ACV/ED		0.001	. . .	2.589	10.795	304	10.85	0.41
000007+2014.3	215155	0.028755	20.237385	Semireg PV	0.951	1.475	MISC		0.005	. . .	213.732	9.064	233	9.07	0.558
000017+2636.4	215156	0.068685	26.608939	Semireg PV	0.418	6.092	MIRA		0.223	. . .	186.65	8.126	119	10.83	1.149
000018+0919.4	215157	0.075405	9.323315	RV Tauri	0.217	10.111	MISC		0.008	. . .	43.541	10.755	239	10.52	0.108
000030−3937.5	215158	0.129525	−39.630347	Beta Persei	0.657	3.202	ED		0.002	. . .	2.553	9.641	410	10.79	0.614
000036+2639.8	215159	0.14772	26.663685	RR Lyrae FM	0.948	0.647	RRAB		0	. . .	0.566	6.65	87	12.61	0.578
000053−1717.5	215160	0.220095	−17.291745	W Ursae Maj	0.893	1.392	ESD/EC		0	. . .	0.298	13.769	345	12.66	0.524
000058+0236.7	215161	0.23553	2.611892	W Ursae Maj	0.938	1.02	EC/DSCT/ESD		0	. . .	0.318	11.915	296	12.94	0.504
000058+1236.5	215162	0.242265	12.607833	Semireg PV	0.218	12.514	MISC		0.023	. . .	5.204	7.917	237	12.64	0.348
000108−3330.1	215163	0.28191	−33.500881	W Ursae Maj	0.959	0.616	EC		0	. . .	0.467	16.315	425	11.51	0.336
000112+0904.7	215164	0.297765	9.078171	Delta Scuti	0.595	2.584	ESD		0	. . .	0.241	11.002	240	10.28	0.08
000116−6037.0	215165	0.316515	−60.615788	Delta Scuti	0.861	2.704	DSCT		0	. . .	0.122	15.126	536	10.03	0.356
000118−3551.7	215166	0.32487	−35.860717	SARG A	0.634	2.205	MISC		0.001	. . .	25.488	8.803	451	9.84	0.254
000119−3505.9	215167	0.32922	−35.097789	SARG B	0.884	1.347	MISC		0	. . .	38.742	7.956	437	10.77	0.258
000120−5834.8	215168	0.334545	−58.580264	LSP	0.649	4.682	MISC		0.011	. . .	375.134	11.847	475	9.53	0.318
000139−0345.4	215169	0.418155	−3.756766	Semireg PV	0.883	2.135	MISC		0.01	. . .	162.98	11.858	344	12.82	0.898
000142−4229.3	215170	0.425685	−42.487419	Semireg PV	0.482	5.452	MISC		0.013	. . .	96.313	8.014	424	10.71	0.379
000147−5714.5	215171	0.44919	−57.242031	Delta Scuti	0.803	3.854	ESD/EC		0	. . .	0.235	17.212	475	11.03	0.15
000155−6707.7	215172	0.475245	−67.130487	RV Tauri	0.268	8.524	MISC		0.085	. . .	1.006	6.464	217	12.74	1.335
000157−5250.1	215173	0.48729	−52.835239	Semireg PV	0.362	4.319	MISC		0.027	. . .	2109.092	11.879	532	10.78	0.419
000158+1357.6	215174	0.49407	13.959879	SARG B	0.338	5.25	MISC		0.006	. . .	74.32	7.228	191	11.52	0.16
000202−6653.3	215175	0.498525	−66.882596	W Ursae Maj	0.976	0.786	EC		0	. . .	0.327	16.748	462	12.16	0.501
000208−1440.5	215176	0.532755	−14.674641	Mira	0.99	0.195	MIRA		0.99	. . .	351.44	15.273	358	8.35	3.534
000221−2929.6	215177	0.582285	−29.493562	W Ursae Maj	0.179	9	ESD/ED		0.005	. . .	3.143	13.906	448	12.32	0.462
000222+0429.6	215178	0.58593	4.494017	SARG A	0.207	6.812	MISC		0.001	. . .	1.002	7.689	302	9.79	0.134
000229−5653.9	215179	0.61572	−56.898676	Weak-line T Tauri	0.213	8.346	ESD/EC/ELL/SR		0.001	. . .	13.223	13.931	474	10.15	0.124
000239−1926.7	215180	0.668595	−19.442961	SARG B	0.411	3.587	MISC		0.007	. . .	453.892	7.716	442	9.69	0.346
000248−2456.7	215181	0.70044	−24.945325	RR Lyrae FM	0.995	0.16	RRAB	RR Lyrae FM	0	. . .	0.493	16.031	387	9.9	0.704
000301−7041.5	215182	0.751575	−70.685852	RR Lyrae FM	0.983	0.314	RRAB		0	. . .	0.554	13.126	340	13.27	0.838

Notes.

[a] ID from the online database http://dotastro.org/.

[b] In decimal degrees.

[c] Statistical significance of the period against a null hypothesis of white noise, in number of σ.

[d] Average V-band magnitude.

[e] Peak to peak amplitude (95th minus 5th quantile).

(This table is available in its entirety in a machine-readable form in the online journal. A portion is shown here for guidance regarding its form and content.)

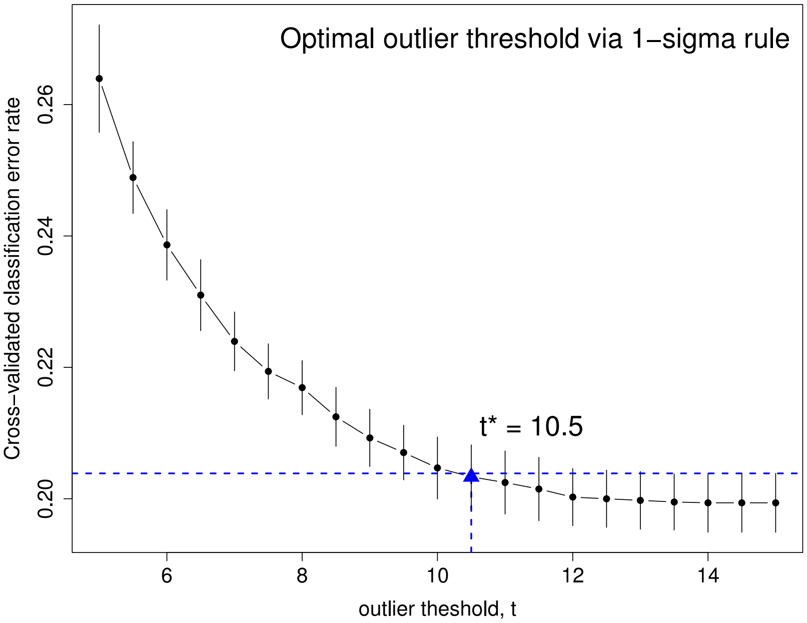


Figure 11. Determination of the optimal anomaly score threshold via cross-validation. As the outlier threshold, t, is reduced, more objects are considered anomalies, and the cross-validated error rate increases (outliers are, by construction, assigned no label, incurring a classification penalty of 1). Using the 1σ rule, which chooses the smallest threshold for which the error rate is within one standard deviation of the default model with no thresholding, we find that the optimal threshold on the anomaly score is $t^* = 10.5$. Adopting this threshold for the ASAS data, we discover 1271 outliers.

(A color version of this figure is available in the online journal.)

8. `Train_Class`: if the ASAS object was in the training set, its training class; otherwise blank
9. `Mira,...,W_Ursae_Maj`: posterior class probabilities for all 28 science classes
10. `P,P_signif`: best-fit period (in days) and its statistical significance (in number of σ)
11. `N_epochs`: number of epochs in the ASAS light curve used to classify the object
12. `V,deltaV`: mean ASAS V-band magnitude and ASAS V-band amplitude.

MACC has been constructed to allow for easy querying of objects of a specified science class, simple searching for outliers, and more advanced queries on several attributes. In supplying the posterior class probabilities for each class, the catalog allows each individual researcher to define their own probability threshold when querying objects. For instance, imagine that scientists A and B are both interested in finding Mira variables, but scientist A requires a highly pure sample, while scientist B simply wants the top 3000 Mira candidates, even if a substantial number of these are non-Miras. Then, scientist A could use a strict threshold, selecting all candidates with $\mathbf{P}(\text{Mira}) > 0.9$ (resulting in 2067 very likely Mira candidates), while scientist B would simply grab the 3000 objects with largest $\mathbf{P}(\text{Mira})$ (which, in this case is equivalent to a Mira probability threshold of 0.370).

4.1. Substituting Different Class Priors

All of the posterior class probabilities given in MACC assume that the prior probability of observing an object of class c_j (before observing any data) is given by the proportion of training set objects that are of class c_j (provided in Table 3). By Bayes' Rule, the posterior MACC class probability for class c_j given the features, $\mathbf{x}_i$, for object i, is

$$\mathbf{P}(c_j|\mathbf{x}_i) = \frac{\mathbf{P}(\mathbf{x}_i|c_j)\mathbf{P}_{\text{tr}}(c_j)}{\sum_{k=1}^{28}\mathbf{P}(\mathbf{x}_i|c_k)\mathbf{P}_{\text{tr}}(c_k)}, \qquad (7)$$

where $\mathbf{P}_{\text{tr}}(c_j)$ is the prior class probability given by the proportion of objects of class c_j in the training set. To exchange a different vector of prior class probabilities, one must multiply each posterior probability from the catalog by the ratio of the new prior to the training set prior and multiply by the corresponding ratio of denominators from Equation (7). For a new prior $\mathbf{P}_{\text{new}}(c_j)$, the new posterior probabilities are given by

$$\mathbf{P}_{\text{new}}(c_j|\mathbf{x}_i) = \mathbf{P}(c_j|\mathbf{x}_i)\frac{\mathbf{P}_{\text{new}}(c_j)}{\mathbf{P}_{\text{tr}}(c_j)}\frac{\sum_{k=1}^{28}\mathbf{P}(\mathbf{x}_i|c_k)\mathbf{P}_{\text{tr}}(c_k)}{\sum_{k=1}^{28}\mathbf{P}(\mathbf{x}_i|c_k)\mathbf{P}_{\text{new}}(c_k)}. \qquad (8)$$

For modified priors, $\mathbf{P}_{\text{new}}$, which are not too dissimilar from $\mathbf{P}_{\text{tr}}$, the last term in Equation (8) will typically be near unity, and thus a reasonable approximation of the modified posterior probabilities can be computed by multiplying the original posteriors by the prior ratio and appropriately re-normalizing. For very dissimilar priors, accurate estimates of all the class-wise densities, $\mathbf{P}(\mathbf{x}_i|c_k)$, would have to be computed and stored on a fine grid of the 72 dimensional feature space, which is both statistically and computationally infeasible.[16] Our recommendation is to only update the posteriors by assuming that the last term in Equation (8) is unity in cases where the prior and posterior class probabilities are all within $\lesssim 0.1$; otherwise the RF classifier should be re-trained with the new prior weights.[17]

The construction in Equation (8) allows us to also condition on additional information such as galactic coordinates (ℓ, b),

[16] Consider the most naive density estimate, a histogram. Constructing a 72-dimensional histogram for each class by binning each feature into 10 bins requires 28×10^{72} values to be computed and stored. Statistically, such a density estimate is unreliable, as the amount of training data is microscopic compared to the vast feature space occupied by 72 dimensions, rendering any simple density estimate useless.

[17] Though there are modern manifold learning methods, which attempt to estimate and exploit lower-dimensional, nonlinear structure in high-dimensional feature spaces to make density estimation tractable, these methods are out of the scope of this paper and likely will also suffer from a lack of labeled data.

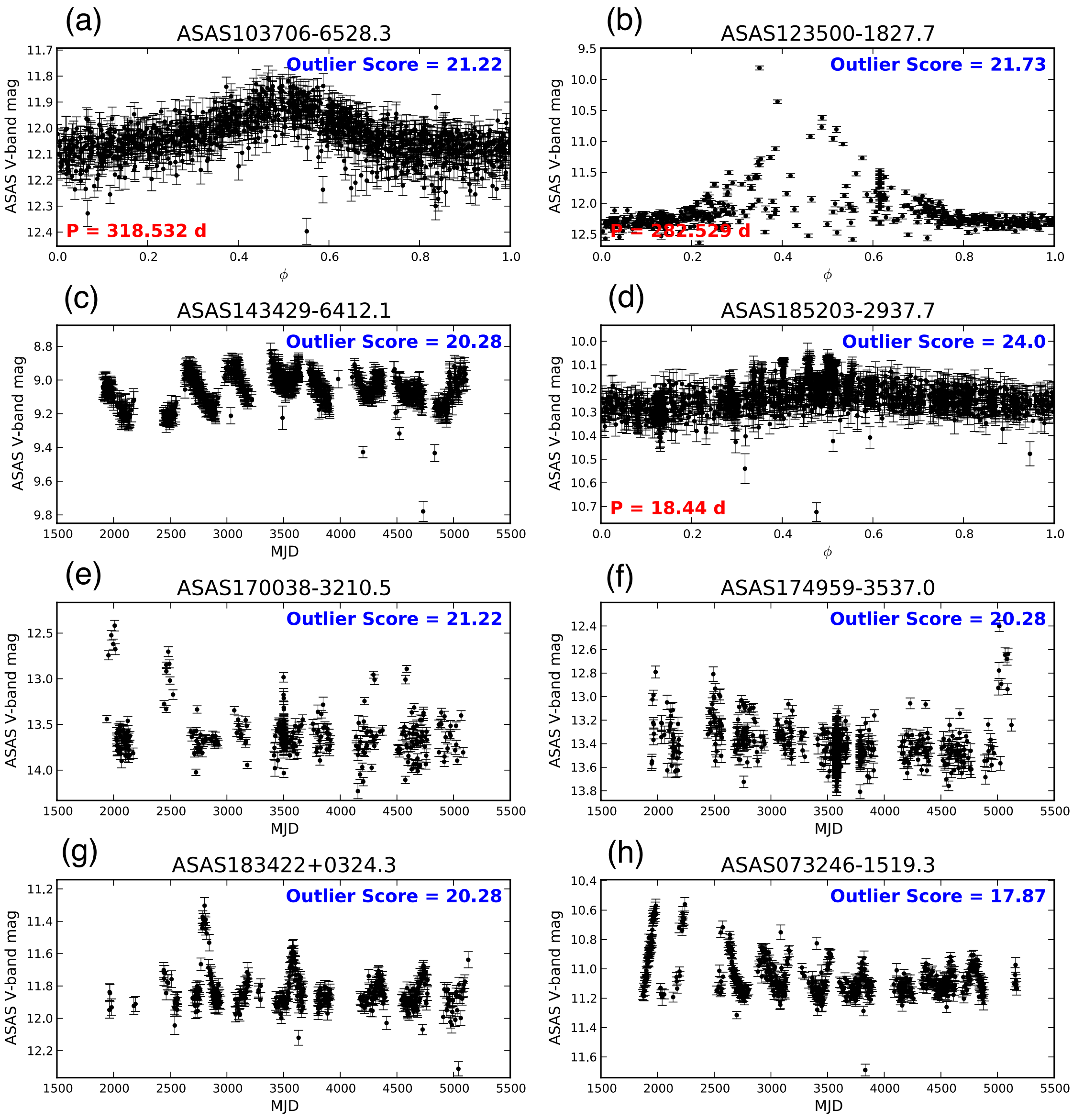


Figure 12. Top outlier light curves as determined from the anomaly score. These light curves are the farthest from their second nearest neighbor in the training set according to the anomaly metric in Equation (6). These sources are outliers because they either occupy an anomalous region of period–amplitude space (a,d), have suppressed high-amplitude variability due to blending with nearby sources (b), show quasi-periodic behavior of irregular type (c,g,h), or have aperiodic variability with >1 mag brightening episodes (e,f). For each of these sources, there are no training instances that capture the observed variability in their ASAS light curves.

(A color version of this figure is available in the online journal.)

median magnitude, and/or distance. For instance, if we have a good theoretical understanding of the expected demographics of variable stars as a function of position in the galaxy, we can imbue that information into the prior probabilities. In other words, before observing any data for a particular object, we can modify its prior class probabilities solely based on its location in the galaxy. This can be a very powerful tool, e.g., for finding star-forming regions near the Galactic plane, where the relative abundance of young stellar objects will be higher (and that information can be inserted into the class prior).

5. COMPARISON TO LITERATURE

We conclude with a comparison of the MACC with a set of papers that have performed classification for ASAS objects. As a first-order validation of MACC classifications, we find that objects with classifications into variable classes, which one would expect to trace the Galactic plane, clearly do. High-probability Classical Cepheids very closely follow the plane, while the Red Giant classes (Mira, SRPV, SARG, RSG, etc.), Chemically Peculiar, Beta Cepheid, and Pulsating Be stars reasonably follow the Plane (sources predominantly have $b < 30°$). Since we do not use any galactic coordinate features in the MACC classification, this serves as an external validation of the quality of the classifier. Other sources that are expected to lie predominately near the Galactic plane (RR Lyrae, WTTS, etc.) are not detected to sufficient distance by ASAS to trace the plane.

In what follows, we continue external verification of MACC by analyzing the similarities and differences between MACC and the popular ACVS catalog. Subsequently, we take a closer look at a handful of papers that have attempted to find, in the

ASAS data, objects of specific subclasses. Overall, we find a high classification agreement rate between MACC and these other works. For the few cases in which the MACC classification disagrees with that of a class-specific paper, the differences can be attributed to poor quality of the ASAS photometry and extra information that was unavailable to our classifier, such as proprietary follow-up data including spectra and high-quality multi-band light curves.

5.1. ASAS Catalog of Variable Stars (ACVS)

As a part of the ACVS, predicted classes are provided for a fraction of the stars. As described in Pojmański (2002), ACVS obtains their classifications using a neural net type algorithm trained on a set of visually labeled ASAS sources, confirmed OGLE Cepheids (Udalski et al. 1999a, 1999b) and OGLE Bulge variable stars (Wozniak et al. 2002). A filter is used to divide strictly periodic from less regular periodic sources. A neural net is trained on the period, amplitude, Fourier coefficients (first four harmonics), $J-H$ and $H - K_s$ colors and IR fluxes to predict the classes of the strictly periodic sources. Many ACVS objects either have multiple labels or are annotated as having low confidence classifications, but no posterior class probabilities are given in the catalog. For less regular periodic sources, location in the $J-H$ versus $H - K_s$ plane is tested; if the object falls within an area of late-type irregular or semi-regular stars, it is assigned the label MISC, else it is inspected by eye. We find that 38,117 ACVS stars, representing 76% of the catalog, are either labeled as MISC, assigned multiple labels, or have low class confidence. The remaining 24% of stars have confident ACVS labels, and provide a set of classifications to compare against our catalog.

In the top panel of Figure 13 we plot the class-wise correspondence between our classifications and the ACVS classes. Overall, there is a 79.7% correspondence between our catalog and ACVS, for the 12,007 sources that are labeled confidently (and not as MISC) in ACVS. For each of the ACVS subclasses, except Population II Cepheid and Multi-Mode Cepheid, we agree on at least 59% of objects. The large disagreement with the Population II and Multi-Mode Cepheids is consistent with the results of Schmidt et al. (2009) who find extreme biases in Cepheid classifications for ACVS. Of 282 stars labeled as Cepheid by ACVS, only 14 were found by Schmidt et al. (2009) to be likely Pop II Cepheids, while all but ∼60 suffered from uncertain period estimates, and ∼50 were rejected as obvious non-Cepheids. We also find that our classifications of First Overtone RR Lyrae, Delta Scuti, and W Ursae Majoris show a significant amount of discrepancy with those of ACVS. In particular, our classifier finds that ∼22% of the stars that ACVS classifies as RRc or Delta Scuti are truly WUMa eclipsing variables.

In the bottom panel of Figure 13, we plot the class-wise correspondence for all 23,209 ASAS sources with MACC outlier score smaller than 3.0. For these more confidently classified objects, MACC has a closer correspondence with ACVS (91.4% for the 8303 objects with confident ACVS class), but still shows high level of disagreement for the non-Classical Cepheids. Of these sources, we find a 98% agreement on Miras, 85% on Classical Cepheids, 99% on RR Lyrae, FM, perfect agreement on 39 Chemically Peculiar stars, 97% on Beta Persei, and 92% on W Uma Majoris sources.

5.1.1. Confident MACC Classifications Missed by ACVS

In addition to having >80% correspondence with ACVS for objects which they confidently label, our MACC catalog identifies many confidently classified sources—having posterior class probability of at least 0.9 for a single class—whose ACVS classification is either uncertain (denoted by a ":" in the catalog) or split between multiple classes. In all, MACC identifies 187 Mira, 22 Classical Cepheid, 122 Fundamental Mode RR Lyrae, 11 First Overtone RR Lyrae, 14 Beta Cephei, 43 Chemically Peculiar, 152 Beta Persei, 210 Beta Lyrae, and 1548 W Uma Majoris candidates that were not found by ACVS. Lowering the confidence threshold from 0.9 to 0.8 yields about 50% more good candidates.

In Figures 14 and 15 we plot, for eight different science classes, the ASAS light curves of selected MACC sources with large class probabilities but whose ACVS classification is different or unconfident. Within each of these classes, the light curves appear as expected for each class of variability. MACC is better able to discover the classes of objects near the magnitude limit of ASAS and whose light curves are of lower signal-to-noise ratio.

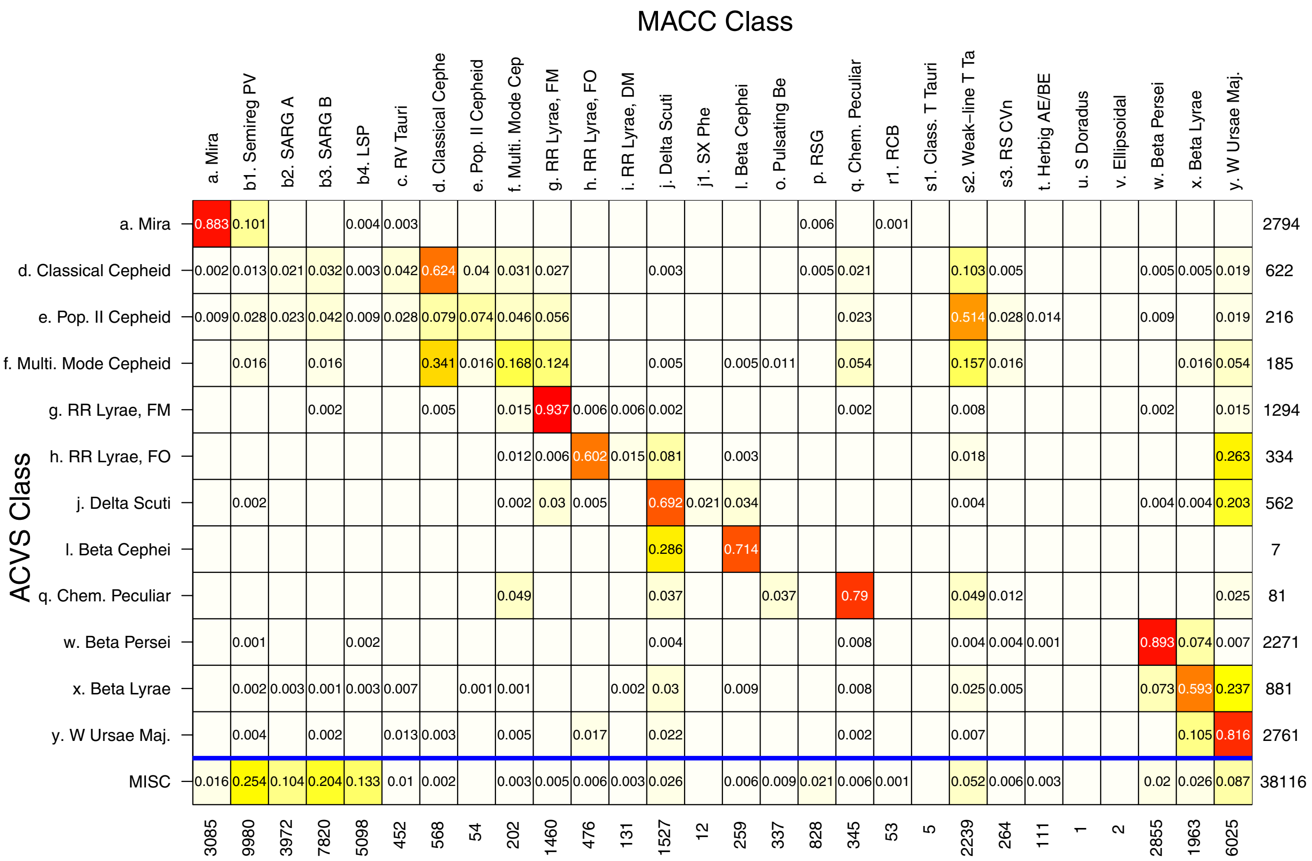
5.2. Classical Cepheids: Berdnikov et al. (2011)

Berdnikov et al. (2011) present multi-band light curves of 49 Classical Cepheid candidates from the ACVS catalog, with data from the 76 cm telescope of the South African Astronomical Observatory and the 40 cm telescope of the Cerro Armazones Observatory of the Catholic University of the North, Chile. From these observations, they are able to confirm that 48 are Classical Cepheids and one, ASAS 100914−5714.6, is a Double-Mode Cepheid. Our classifier correctly identifies 46 of these 48 Classical Cepheids. See Table 5 for a complete listing of our catalog classification, posterior probability of Classical Cepheid, ranking of Classical Cepheid probability out of all 50 K ASAS sources, and anomaly score for all 49 objects observed by Berdnikov et al. (2011). None of these objects is in our MACC training set.

For two of these sources, ASAS 073453−2651.3 and ASAS 075750−2923.5, our catalog identifies the objects as non-Classical Cepheids. For the former, the object has a very high anomaly score of 15.13, meaning that its classification as WTTS should not be trusted. This object has a median ASAS magnitude around 14, which is near the ASAS detection limit, thus its light curve is noisy and contains many non-detections when the object is dimmer than median. The non-detections give the source a depressed amplitude from what is expected for a Classical Cepheid, and thus the source is flagged as anomalous. For the latter, which has a period of 2.586 days, there is significant scatter in the phased ASAS light curve and a relatively low amplitude, making its ASAS light curve more consistent with a Multi-Mode Cepheid. It is likely the presence of a bright neighbor to ASAS 075750−2923.5 that causes this scatter and depressed amplitude. However, the light curve of Berdnikov et al. (2011), which only contains nine epochs of data, does not completely rule out a Multi-Mode pulsator.

5.3. Beta Cephei: Pigulski (2005)

In the work of Pigulski (2005), 14 new Beta Cephei stars appearing in ACVS were confirmed (in addition to 4 other previously known Beta Cephei stars). Starting with all 37 stars whose ACVS classification includes BCEP as a possible class, the author makes selection cuts based on the ASAS periodogram and any available multi-band photometry and/or spectral type, finding 14 stars that the author deems as unambiguous. Then, with a broader set of 1700 ASAS stars, Pigulski detects four more bona fide candidates using the same selection criteria.



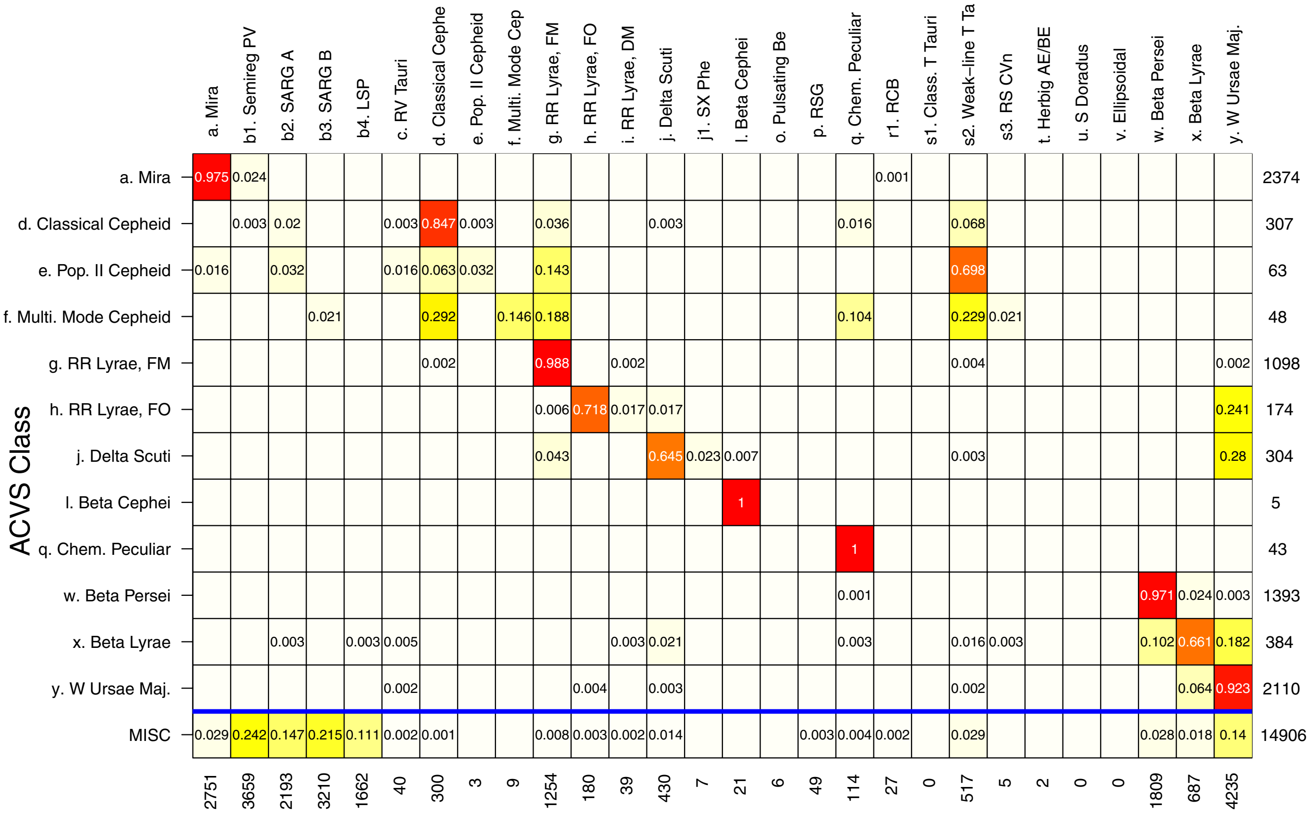


Figure 13. Top: correspondence of the MACC to the ACVS classifications for all 50,124 sources. Rows are normalized to sum to 100%. Marginal counts are listed to the right and bottom of the table. There is a 79.7% total correspondence between our classifications and the ACVS labels for the 12,007 objects whose ACVS classification is a single confident class not equal to MISC. Bottom: same for the subset of 23,209 ASAS sources with outlier score smaller than 3. The agreement rate between MACC and ACVS for the subset of these sources with confident ACVS class (8303 objects) is 91.4%.

(A color version of this figure is available in the online journal.)

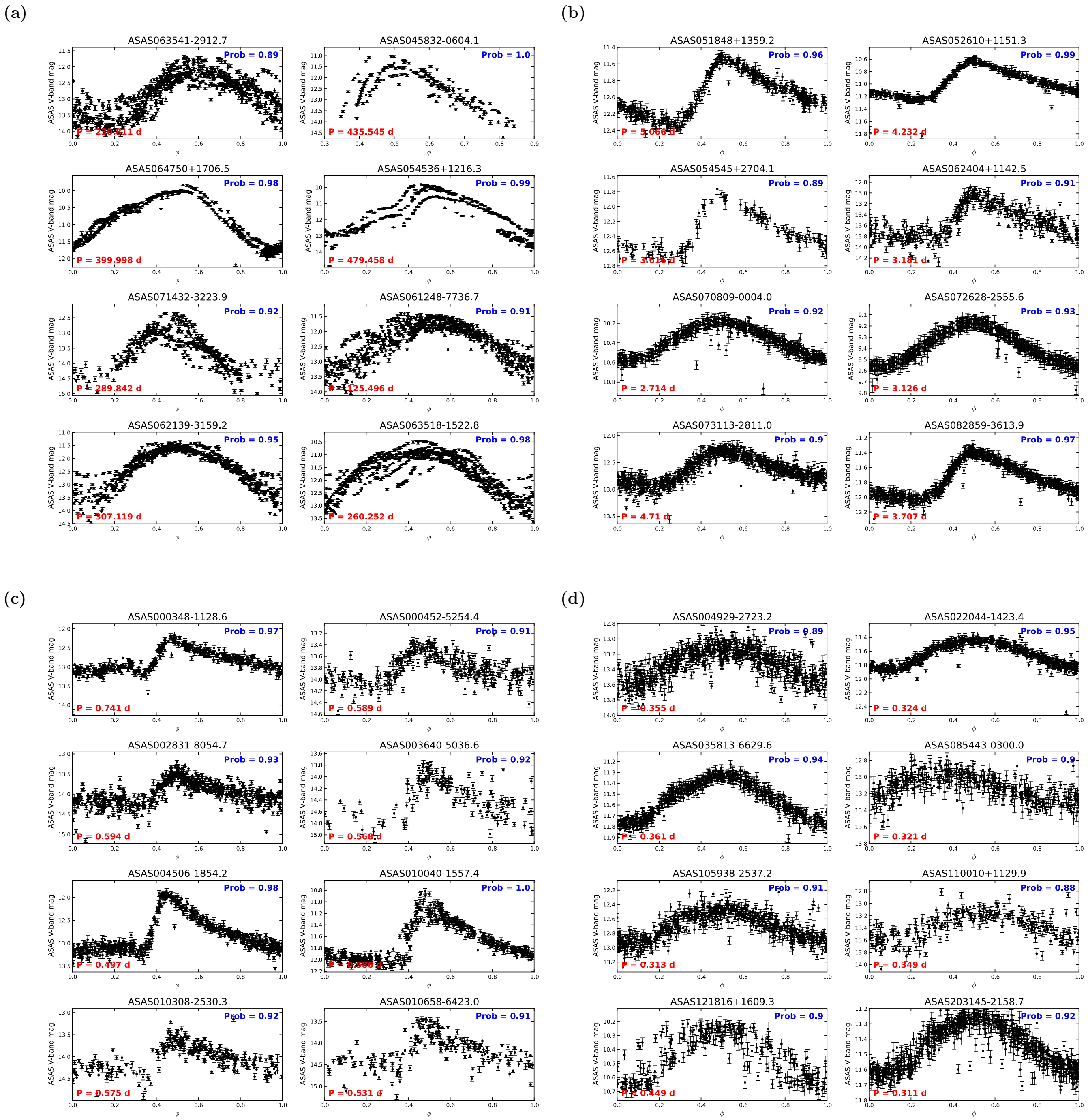


Figure 14. ASAS light curves for arbitrarily chosen candidates with a high probability of being (a) Mira, (b) Classical Cepheid, (c) Fundamental Mode RR Lyrae, and (d) First Overtone RR Lyrae whose ACVS classification either includes multiple classes, is insecure or MISC, or otherwise differs from that of MACC.

(A color version of this figure is available in the online journal.)

In Table 6 we report our catalog's classification for each of the Pigulski (2005) Beta Cephei. Note that all but one of these sources was included in the MACC training set. We misidentify as a Delta Scuti star the one object (ASAS 161858−5103.5) that was not included in the training set. This star is located directly in the Galactic plane with a Galactic latitude of $-0\overset{\circ}{.}536$, and suffers from heavy extinction. Thus its observed colors are more typical of the comparatively redder class of Delta Scuti stars than the bluer class of Beta Cephei. With a Beta Cephei posterior class probability of 0.17, it ranks within the top 500 Beta Cephei candidates.

5.4. Double-Mode RR Lyrae: Szczygieł & Fabrycky (2007)

Szczygieł & Fabrycky (2007) perform a search for multiple-pulsating RR Lyrae stars in ASAS. Starting with all objects with an RR Lyrae classification in ACVS, this study first culled out obvious non-RR Lyrae stars via visual inspection. They prewhiten each ASAS RR Lyrae light curve at the pulsation period

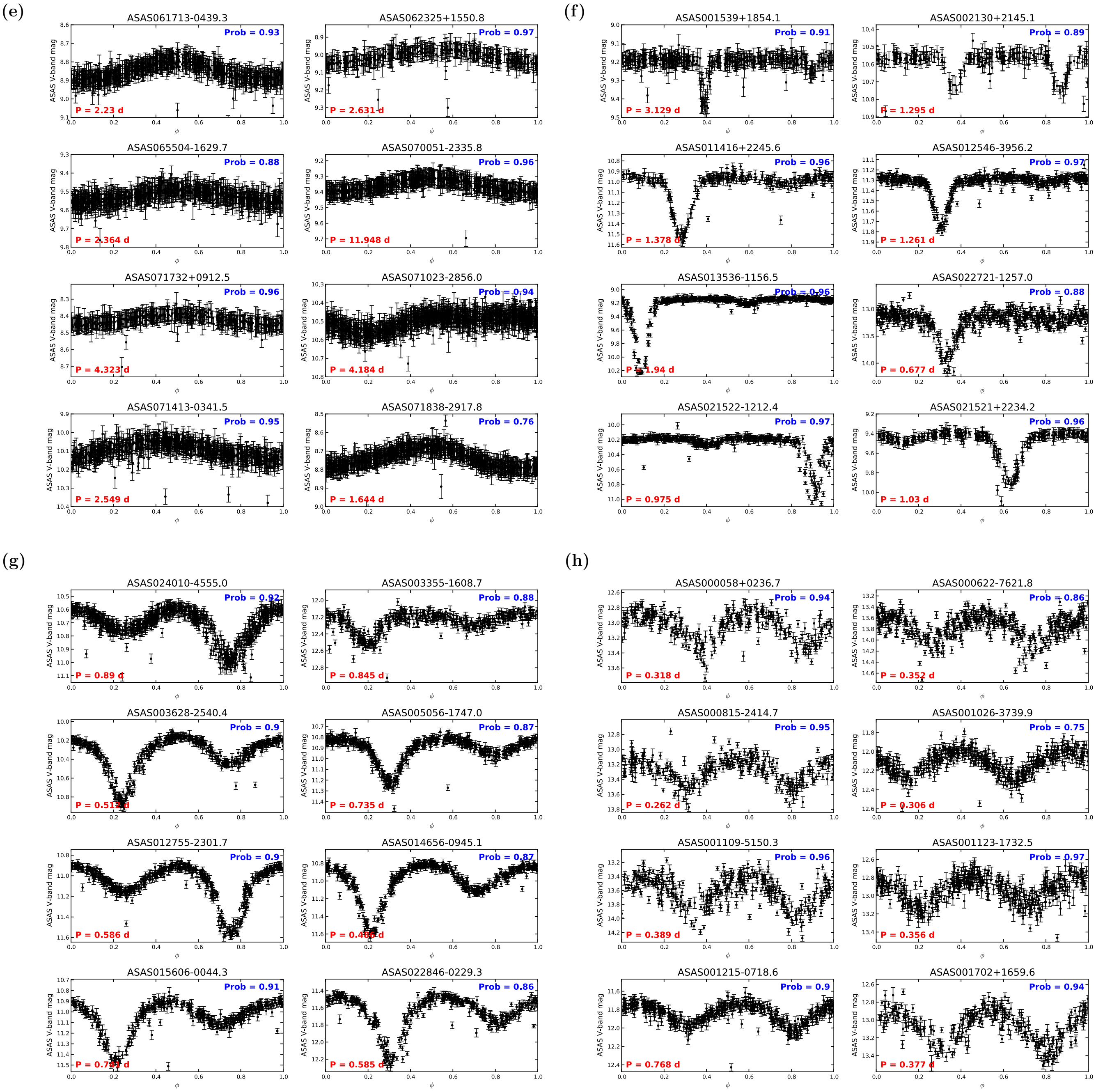


Figure 15. Same as Figure 14 for: (e) Chemically Peculiar, (f) Beta Persei, (g) Beta Lyrae, and (h) W Uma Majoris.

(A color version of this figure is available in the online journal.)

and run the CLEAN algorithm to find any significant periodicity in the residual light curves. From this analysis, they identify of order 150 Blazhko affected RR Lyrae and 19 Double-Mode RR Lyrae stars. The Double-Mode pulsators were identified by making cuts on the pulsation period (P_0) and the ratio of the overtone to fundamental periods ($0.735 \leqslant P_1/P_0 \leqslant 0.755$) and confirmed via visual inspection.

The MACC classification, posterior probability of Double-Mode RR Lyrae, ranking of RRd amongst all ASAS sources, and anomaly score for the 19 confirmed RRd from Szczygieł & Fabrycky (2007) are in Table 7. MACC correctly classifies all 19 stars even though only two of them were in our training set. Each of the stars has posterior probability of being a Double-Mode RR Lyrae of >0.65 and each ranks within the top 29 RRd candidates.

5.5. *Orion Belt Objects: Caballero et al. (2010)*

In a search for high-amplitude variable stars in the Orion Belt, Caballero et al. (2010) identify 32 variable stars from ASAS photometry, proper motions, and infrared photometry (2MASS and the *IRAS*). They perform an extensive literature search on these objects and visual analysis to determine a likely classification for each. Of these 32 variable stars, 13 are in our catalog, and their classifications are listed in Table 8. Our classifications agree with those of Caballero et al. (2010) for 9 of the 13 objects.

Table 5
Classification Catalog Results for Classical Cepheid Stars Confirmed by Berdnikov et al. (2011)

ASAS ID	Predicted Class	P(Classical Cepheid)	Rank CCeph	Anomaly Score	In Training
052610+1151.3	Classical Cepheid	0.993	53	0.23	No
052706+1656.2	Classical Cepheid	0.893	178	1.17	No
062939−1840.5	Classical Cepheid	0.873	202	1.27	No
064037+1143.6	Classical Cepheid	0.923	140	0.87	No
064829−1014.2	Classical Cepheid	0.858	218	1.38	No
070355−1752.8	Classical Cepheid	0.937	125	0.63	No
071342−1737.2	Classical Cepheid	0.813	262	2.18	No
073113−2811.0	Classical Cepheid	0.9	170	1.76	No
073453−2651.3	Weak-line T Tauri	0.166	604	15.13	No
073502−3554.9	Classical Cepheid	0.843	242	1.65	No
074925−3814.4	Classical Cepheid	0.906	161	1.1	No
075345−3658.2	Classical Cepheid	0.995	45	0.2	No
075358−2822.1	Classical Cepheid	0.812	264	1.9	No
075750−2923.5	Multi-Mode Cepheid	0.125	683	4.03	No
075840−3330.2	Classical Cepheid	0.944	118	0.65	No
075912−2641.9	Classical Cepheid	0.703	320	4.81	No
080500−2851.8	Classical Cepheid	0.937	126	0.56	No
080511−3421.7	Classical Cepheid	0.921	142	0.9	No
080927−3315.7	Classical Cepheid	0.898	173	1.23	No
081025−3828.4	Classical Cepheid	0.872	205	1.53	No
081026−3231.3	Classical Cepheid	0.931	130	0.91	No
082117−3845.3	Classical Cepheid	0.817	259	1.42	No
082127−3825.3	Classical Cepheid	0.896	176	1.33	No
082859−3613.9	Classical Cepheid	0.968	86	0.42	No
083130−4429.3	Classical Cepheid	0.636	349	5.41	No
083426−3559.1	Classical Cepheid	0.948	115	0.59	No
083611−3903.7	Classical Cepheid	0.823	257	2.13	No
084127−4353.6	Classical Cepheid	0.832	249	2.46	No
090436−4633.2	Classical Cepheid	0.627	355	8.35	No
090932−5359.3	Classical Cepheid	0.909	158	1.16	No
092758−5218.9	Classical Cepheid	0.848	231	1.78	No
093005−5137.5	Classical Cepheid	0.871	207	1	No
094819−5748.6	Classical Cepheid	0.771	284	3.31	No
094827−5801.1	Classical Cepheid	0.957	105	0.48	No
100914−5714.6	Classical Cepheid	0.637	347	9.53	No
101037−5817.8	Classical Cepheid	0.779	281	2.91	No
101538−5933.1	Classical Cepheid	0.336	502	5.76	No
103627−6211.6	Classical Cepheid	0.814	261	1.75	No
112039−6149.9	Classical Cepheid	0.826	253	2.86	No
115701−6218.7	Classical Cepheid	0.284	534	8.26	No
122240−6209.5	Classical Cepheid	0.844	238	1.75	No
123804−3831.4	Classical Cepheid	0.902	166	1.1	No
140742−6315.4	Classical Cepheid	0.843	240	1.9	No
150547−5823.0	Classical Cepheid	0.883	194	1.23	No
152021−5807.3	Classical Cepheid	0.814	260	3.26	No
164120−4739.6	Classical Cepheid	0.328	505	4.21	No
173253−3554.7	Classical Cepheid	0.743	304	4.24	No
174134−4854.6	Classical Cepheid	0.609	366	4.81	No
181416−0920.4	Classical Cepheid	0.632	351	3.69	No

For four objects, we disagree with the classifications of Caballero et al. (2010). The star ASAS 053621−0210.9 (PQ Ori) was found by us to be a semi-detached (Beta Lyrae) eclipsing system, while Caballero et al. (2010) note that although it has been identified as a possible young stellar object in the literature, its colors are too blue and it is more likely a field star. The star ASAS 053946−0055.9 was identified by us as either an LSP or RS CVn, consistent with the classification of Schirmer et al. (2009), while Caballero et al. (2010) retain it as an uncertain T Tauri candidate. The star ASAS 053543−0034.6 is claimed by Caballero et al. (2010) to have signs of youth; however, we find significant periodicity on 86.61 day timescales, which is consistent with the pulsations of an RV Tauri star. Finally, ASAS 053642+0038.5 is identified by our catalog as a likely W Ursae Majoris candidate due to its tell-tale eclipsing structure on 1.06 day timescales; Caballero et al. (2010) claim that it is a possible HAeBe star, though they note that it has anomalous brightness.

Table 6
Classification Catalog Results for Beta Cephei Stars in Pigulski (2005)

ASAS ID	Predicted Class	P(Beta Cephei)	Rank BetCep	Anomaly Score	In Training
091731−5250.3	Beta Cephei	0.921	5	2.73	Yes
180233−4005.2	Beta Cephei	0.972	1	0.96	Yes
191715+0103.6	Beta Cephei	0.909	9	2.8	Yes
212329+0955.9	Beta Cephei	0.884	14	5.8	Yes
122213−6320.8	Beta Cephei	0.916	6	2.38	Yes
150955−6530.4	Beta Cephei	0.943	3	1.8	Yes
161858−5103.5	Delta Scuti	0.17	480	3.41	No
164409−4719.1	Beta Cephei	0.857	16	2.51	Yes
164630−4701.2	Beta Cephei	0.837	19	1.88	Yes
164939−4431.7	Beta Cephei	0.829	20	2.58	Yes
165314−4345.0	Beta Cephei	0.913	7	2.28	Yes
165554−4808.8	Beta Cephei	0.872	15	1.92	Yes
171218−3306.1	Beta Cephei	0.903	10	2.28	Yes
180808−3434.5	Beta Cephei	0.896	12	2.48	Yes
181716−1527.1	Beta Cephei	0.911	8	2.48	Yes
182610−1704.3	Beta Cephei	0.901	11	1.88	Yes
182617−1515.7	Beta Cephei	0.807	21	4.13	Yes
182726−1442.1	Beta Cephei	0.895	13	1.99	Yes

Table 7
Classification Catalog Results for Double-Mode RR Lyrae Stars in Szczygieł & Fabrycky (2007)

ASAS ID	Predicted Class	P(RRd)	Rank RRd	Anomaly Score	In Training
032820−6458.7	RR Lyrae DM	0.834	11	2.03	No
040054−4923.8	RR Lyrae DM	0.886	4	1.64	No
081610−6644.8	RR Lyrae DM	0.778	20	2.85	No
084747−0339.1	RR Lyrae DM	0.759	21	4.41	No
122509−2139.9	RR Lyrae DM	0.658	28	3.83	No
133439+2416.6	RR Lyrae DM	0.825	12	2.18	No
141539+0010.1	RR Lyrae DM	0.707	26	3.65	No
151735−0105.3	RR Lyrae DM	0.723	22	3.42	No
173726+1122.4	RR Lyrae DM	0.85	9	2.18	No
183952−3200.9	RR Lyrae DM	0.653	29	5.54	Yes
184035−5350.7	RR Lyrae DM	0.933	1	0.9	No
193933−6528.9	RR Lyrae DM	0.863	6	2.32	No
195612−5043.7	RR Lyrae DM	0.925	2	1.06	Yes
210726+0110.3	RR Lyrae DM	0.835	10	2.36	No
211848−3430.4	RR Lyrae DM	0.79	18	3.39	No
212721−1908.0	RR Lyrae DM	0.859	7	1.79	No
213437−4907.5	RR Lyrae DM	0.783	19	3.12	No
230449−3345.3	RR Lyrae DM	0.814	14	2.91	No
235622−5329.4	RR Lyrae DM	0.824	13	2.36	No

Table 8
Classification Catalog Results for Orion Belt Variables in Caballero et al. (2010)

ASAS ID	Predicted Class	P(Class)	Anomaly Score	In Training	Caballero Class
054354−0243.6	W Ursae Maj	0.998	0.3	No	Contactbinary
053848−0227.2	Weak-line T Tauri	0.371	9	No	TTauri
053621−0210.9	Beta Lyrae	0.849	3.24	No	HAeBe
053739−0146.3	Mira	0.995	0.12	No	Giant
053757−0140.8	Semireg PV	0.874	2.41	No	Giant
053126−0058.6	W Ursae Maj	0.651	3.2	No	Unknown
053946−0055.9	LSP	0.332	9.64	No	TTauri?
052725−0035.2	SARG B	0.582	2.32	No	Giant
053543−0034.6	RV Tauri	0.62	4.15	No	TTauri
052634−0019.5	SARG B	0.856	2.85	No	Giant
054612+0032.4	RV Tauri	0.301	5.67	No	Unknown
053642+0038.5	W Ursae Maj	0.503	7.93	No	HAeBe?
053348+0055.6	SARG B	0.403	3.57	No	Giant

6. CONCLUSIONS

We have presented an end to end methodology for creating a probabilistic classification catalog for a time-domain survey of variability. With growing data volumes and rates, these types of automated classification catalogs become necessary for astronomers to make sense of such a vast amount of data and to optimize the allocation of limited follow-up resources. Though the machine-learned construction of accurate classification catalogs is certainly a difficult undertaking, we have shown that sub-20% error rates are achievable even with as many as 28 classes and subclasses of stellar variability. Furthermore, we have motivated the importance of disseminating probabilistic classifications with full disclosure of class priors, allowing each user freedom to trade class purity for efficiency and to use full probability vectors in performing astrophysical inference (for a recent use of probabilities for cosmological parameter estimation, see Newling et al. 2012). Additionally, it is crucial that the classification probabilities be calibrated so that the natural interpretation of probability holds, allowing for faithful propagation of that information to downstream analyses.

As a test case for the methodologies presented in this paper, and those adopted from Richards et al. (2011, 2012), we build and make publicly available a 28-class MACC of 50,124 sources that are included in the ACVS. We show that accurate classifications are possible for such a complex, noisy, and diverse data set of photometric light curves. Furthermore, we demonstrate that calibrated probabilities are attainable using straightforward methodology and that semi-supervised anomaly detection can discover interesting objects that do not fit within a predefined classification taxonomy. Comparisons of our MACC with existing ASAS classifications, including those in ACVS, are favorable and we eagerly await more intense scrutiny of the publicly available MACC from the astronomical community. Inevitably many of our top classifications will be proven incorrect, but that is expected by the very nature of the product: it is, instead, the testing of the aggregate accuracy of our probabilistic classifications that are of most long-term interest.

Some degrees of the predicted accuracy and functionality of the MACC catalog have already been demonstrated in the concurrently submitted paper of Miller et al. (2012). In that paper, MACC was used to search for previously unknown R Coronae Borealis and DY Persei stars in ASAS. Their search through the top MACC RCB candidates yielded 12 likely RCB/DYPers stars, whereby they confirmed with new and archival spectroscopic observations the discovery of four RCB stars and four DYPers, increasing the number of known Galactic DYPers from 2 to 6. Miller et al. (2012) demonstrate that the MACC catalog recovers ASAS candidates that would have been missed via the typical search method which uses hard cuts on the amplitude and periodicity of the light curves, and that a prohibitive number of objects would have to be manually searched via those traditional methods to recover all of the newly discovered objects. This is powerful validation that machine-learned probabilistic classification can facilitate astronomical discovery and enable scientific results.

Moving forward, there remain many pending tasks for our machine-learned approach to classification catalogs. First, we have not touched on the question of discovery of variability, only on classification once variable objects have been identified. Recently, Shin et al. (2009) have introduced a machine-learning approach to variability selection which we will expand to develop new procedures. Second, the size and scope of MACC, at 50k variable stars at a brightness level reaching 14th magnitude, is rather small and limited. Tackling larger catalogs with millions of sources will test the feasibility of our algorithms and robustness of our statistical approaches. Third, the future of time-domain surveys is multi-band light curves (e.g., DES, LSST). Neglecting the full use and exploitation of multi-band photometry would mean throwing away much useful information. Last, a large component of the catalog-building techniques that we have presented is the constant feedback from the automated classifier and the astronomical community. From compiling large and representative training sets to inventing new features that probe different types of variability, constant injection of more information into the machine learner is essential to optimize the accuracy, information gain, and ultimately the scientific impact of the catalog.

The authors acknowledge the generous support of a CDI grant (no. 0941742) from the National Science Foundation. This work was performed in the CDI-sponsored Center for Time Domain Informatics (http://cftd.info). We also thank IBM and CITRIS for providing the 280 core cluster at Berkeley, which was used to perform feature computations and classifier evaluations. We are very grateful to G. Pojmański for building and publicly releasing the ACVS.

APPENDIX

MODIFYING PARASITE FREQUENCIES

In a previous version of the manuscript and catalog, we employed a method to treat parasite frequencies. Parasite frequencies are caused by genuine variability in photometry and are *not* due to astrophysical phenomena. In ground-based observations these parasitic frequencies typically occur at 1, 2, 3, etc., cycles per day due to the rotational period of Earth and inadequate treatment of atmospheric extinction. Indeed, we see prominent overdensities at each of these frequencies in the ASAS data. Originally, we intended to correct the period estimates for those (many) sources whose first Lomb–Scargle frequency occurs within ±0.05 cycles/day of a parasite frequency. Our supposition was that the plethora of sources at period of 1, 1/2, 1/3, $\cdots$ day would confuse the classifier, resulting in worse performance and debilitating artifacts (such as all sources with 1 day period being classified as arising from the same variability class). To avoid this scenario, we used the 1/freq1_harmonics_freq_0–freq_signif plane to discriminate likely cases of parasite frequencies from stars whose period was attributed to true astrophysical variability. We achieved this using an admittedly ad hoc iterative procedure of fitting a separating curve in 1/freq1_harmonics_freq_0–freq_signif space, manually inspecting ambiguous cases near the border and refitting the separation boundary. For each star which was deemed to have a parasite frequency, we threw away that frequency estimate and instead used the next pre-whitened (non-parasitic) Lomb–Scargle frequency estimate. However, given the doubts of the referee as to the legitimacy of the procedure by which we modified the periods for these sources, we decided to analyze the sensitivity of the classification results to the parasitic frequencies. To do this, we ran a comparison of the RF classifications using the following two feature sets: (1) the original feature set (without modified Lomb–Scargle frequencies) and (2) the feature set with modified Lomb–Scargle frequencies via the prescription outlined above. The results showed that the classifications on the set of 50,124 ACVS sources using these two feature sets were very similar. The classifications from

Table 9
Classification Breakdown of Objects with 1 day LS Periods without and with Correction for Parasite Frequency

Class	No Correction	Parasite Correction	Difference
Mira	45	50	+5
SRPV	2455	2019	−436
SARG A	520	546	+26
SARG B	1382	1756	+374
LSP	37	289	+252
RV Tauri	33	34	+1
Classical Cepheid	2	3	+1
Pop. I Cepheid	1	0	−1
Multi. Mode Cepheid	16	3	−13
RR Lyrae, FM	20	18	−2
Delta Scuti	0	22	+22
Pulsating Be	9	7	−2
Chem. Peculiar	18	17	−1
Class. T Tauri	4	6	+2
Weak-line T Tauri	619	404	−215
RS CVn	9	4	−5
Herbig Ae/Be	44	41	−3
Beta Persei	189	182	−7
Beta Lyrae	95	85	−10
W Ursae Maj.	34	22	−12

these two feature sets agreed for 88.0% of all sources, and for 99.2% of all sources (22,936 of 23,129 sources) whose top class had probability over 0.5 with the original feature set. Moreover, classifications actually got slightly better by not modifying the periods of the parasitic sources: the cross-validation error rate improved by ∼1% and agreement with ACVS increased by 0.7% by not modifying the parasite frequencies. Moreover, no significant artifacts occurred due to the abundance of sources with parasite frequencies. Of the 5725 sources consistent with a parasite frequency of 1 cycle per day, the breakdown of classes before and after alteration of the frequency is tabulated in Table 9. It is clear that no single class dominates when no correction for parasite frequencies is performed. There are a few red giant classes and WTTS that changed by more than 200 sources, but all of these differences occurred for low probability (<0.5) sources, and no single class gained undue influence. Thus, we concluded that even if we did not alter the estimated frequencies of parasite sources, the RF classifier was able to learn to ignore these parasite frequencies and instead use other more discriminating features to perform classification. Hence, in the final version of the paper and MACC catalog, no modification of the parasite frequencies was performed.

REFERENCES

Berdnikov, L. N., Kniazev, A. Y., Sefako, R., et al. 2011, Astron. Rep., 55, 816

Bhattacharyya, S., Richards, J. W., Rice, J., et al. 2011, in Proc. Statistical Challenges in Modern Astronomy V, ed. E. D. Feigelson & G. J. Baba (New York: Springer), 483

Blomme, J., Sarro, L. M., O'Donovan, F. T., et al. 2011, MNRAS, 418, 96

Borne, K. D., Laher, R., Ivezic, Z., & Hamam, N. LSST Collaboration. 2009, BAAS, 41, 213

Bostrom, H. 2008, in Proc. 7th International Conf. Machine Learning and Applications, ed. M. A. Wani, X. W. Chen, D. Casasent et al. (San Diego, CA: IEEE Computer Society), 121

Breiman, L. 2001, Mach. learn., 45, 5

Brewer, J. M., Bloom, J. S., Kennedy, R., & Starr, D. L. 2009, in ASP Conf. Ser. 411, Astronomical Data Analysis Software and Systems XVIII, ed. D. A. Bohlender, D. Durand, & P. Dowler (San Francisco, CA: ASP), 357

Brier, G. 1950, Mon. Weather Rev., 78, 1

Caballero, J. A., Cornide, M., & de Castro, E. 2010, Astron. Nachr., 331, 257

Clayton, G. C. 1996, PASP, 108, 225

Craven, P., & Wahba, G. 1979, Numer. Math., 31, 377

Debosscher, J., Sarro, L. M., Aerts, C., et al. 2007, A&A, 475, 1159

Debosscher, J., Sarro, L. M., López, M., et al. 2009, A&A, 506, 519

Dubath, P., Rimoldini, L., Süveges, M., et al. 2011, MNRAS, 414, 2602

Eyer, L., Jan, A., Dubath, P., et al. 2008, in AIP Conf. Proc. 1082, Classification and Discovery in Large Astronomical Surveys, ed. C. A. L. Bailer-Jones (Melville, NY: AIP), 257

Golub, G., Heath, M., & Wahba, G. 1979, Technometrics, 215

Hastie, T., Tibshirani, R., & Friedman, J. 2009, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Vol. 21 (2nd ed.; New York: Springer)

Herbig, G. H., & Bell, K. R. (ed.) 1988, Third Catalog of Emission-Line Stars of the Orion Population : 3 : 1988 (Santa Cruz: Lick Observatory)

Herbst, W., Herbst, D. K., Grossman, E. J., & Weinstein, D. 1994, AJ, 108, 1906

Kim, D.-W., Protopapas, P., Byun, Y.-I., et al. 2011, ApJ, 735, 68

Long, J. P., Karoui, N. E., Rice, J. A., Richards, J. W., & Bloom, J. S. 2012, PASP, 124, 280

Massey, P., & Olsen, K. A. G. 2003, AJ, 126, 2867

Miller, A. A., Richards, J. W., Bloom, J. S., et al. 2012, ApJ, 755, 98

Neugent, K. F., Massey, P., Skiff, B., & Meynet, G. 2012, ApJ, 749, 177

Newling, J., Bassett, B., Hlozek, R., et al. 2012, MNRAS, 421, 913

Pigulski, A. 2005, Acta Astron., 55, 219

Platt, J. 1999, Advances in Large Margin Classifiers, 10 (Boston: MIT Press), 61

Pojmański, G. 1997, Acta Astron., 47, 467

Pojmański, G. 2002, Acta Astron., 52, 397

Pojmański, G. 2003, Acta Astron., 53, 341

Pojmański, G., & Maciejewski, G. 2004, Acta Astron., 54, 153

Pojmański, G., & Maciejewski, G. 2005, Acta Astron., 55, 97

Pojmański, G., Pilecki, B., & Szczygiel, D. 2005, Acta Astron., 55, 275

Protopapas, P., Giammarco, J. M., Faccioli, L., et al. 2006, MNRAS, 369, 677

Rebbapragada, U., Protopapas, P., Brodley, C. E., & Alcock, C. 2009, in ASP Conf. Ser. 411, Astronomical Data Analysis Software and Systems XVIII, ed. D. A. Bohlender, D. Durand, & P. Dowler (San Francisco, CA: ASP), 264

Richards, J. W., Starr, D. L., Brink, H., et al. 2012, ApJ, 744, 192

Richards, J. W., Starr, D. L., Butler, N. R., et al. 2011, ApJ, 733, 10

Robertson, T., Wright, F., Dykstra, R., & Robertson, T. 1988, Order Restricted Statistical Inference, Vol. 229 (New York: Wiley)

Schirmer, J., Bernhard, K., & Lloyd, C. 2009, Open Eur. J. Var. Stars, 105, 1

Schmidt, E. G., Hemen, B., Rogalla, D., & Thacker-Lynn, L. 2009, AJ, 137, 4598

Shin, M., Sekora, M., & Byun, Y. 2009, MNRAS, 400, 1897

Stekhoven, D. J., & Bühlmann, P. 2012, Bioinformatics, 28, 112

Stetson, P. B. 1996, PASP, 108, 851

Strassmeier, K. G., Hall, D. S., Zeilik, M., et al. 1988, A&AS, 72, 291

Strom, K. M., Strom, S. E., Edwards, S., Cabrit, S., & Skrutskie, M. F. 1989, AJ, 97, 1451

Szczygieł, D. M., & Fabrycky, D. C. 2007, MNRAS, 377, 1263

Torres, C. A. O., Quast, G. R., da Silva, L., et al. 2006, A&A, 460, 695

Tyson, J. A. 2002, Proc. SPIE, 4836, 10

Udalski, A., Soszynski, I., Szymanski, M., et al. 1999a, Acta Astron., 49, 223

Udalski, A., Soszynski, I., Szymanski, M., et al. 1999b, Acta Astron., 49, 437

Varón, C., Alzate, C., Suykens, J. A. K., & Debosscher, J. 2011, A&A, 531, A156

Walter, F. M. 1986, ApJ, 306, 573

Wenger, M., Ochsenbein, F., Egret, D., et al. 2000, A&AS, 143, 9

Wozniak, P. R., Udalski, A., Szymanski, M., et al. 2002, Acta Astron., 52, 129

Zacharias, N., Monet, D. G., Levine, S. E., et al. 2004, BAAS, 36, 1418

Zadrozny, B., & Elkan, C. 2001, in International Conference on Machine Learning (ICML), ed. C. E. Brodley & A. Pohoreckyj Danyluk (San Francisco, CA: Morgan Kaufmann), 609

Zechmeister, M., & Kürster, M. 2009, A&A, 496, 577