THE EB FACTORY PROJECT. II. VALIDATION WITH THE KEPLER FIELD IN PREPARATION FOR K2 AND TESS

MAHMOUD PARVIZI1,2, MARTIN PAEGERT3, AND KEIVAN G. STASSUN1,3

1 Department of Physics and Astronomy, Vanderbilt University, VU Station B 1807, Nashville, TN 37235, USA; mahmoud.parviz@vanderbilt.edu
2 Department of Physics and Astronomy, Austin Peay State University, 601 College St. Clarksville, TN 37044, USA
3 Department of Physics, Fisk University, 1000 17th Ave. N., Nashville, TN 37208, USA

ABSTRACT

Large repositories of high precision light curve data, such as the Kepler data set, provide the opportunity to identify astrophysically important eclipsing binary (EB) systems in large quantities. However, the rate of classical “by eye” human analysis restricts complete and efficient mining of EBs from these data using classical techniques. To prepare for mining EBs from the upcoming K2 mission as well as other current missions, we developed an automated end-to-end computational pipeline—the Eclipsing Binary Factory (EBF)—that automatically identifies EBs and classifies them into morphological types. The EBF has been previously tested on ground-based light curves. To assess the performance of the EBF in the context of space-based data, we apply the EBF to the full set of light curves in the Kepler “Q3” Data Release. We compare the EBs identified from this automated approach against the human generated Kepler EB Catalog of ~2600 EBs. When we require EB classification with $>90\%$ confidence, we find that the EBF correctly identifies and classifies eclipsing contact (EC), eclipsing semi-detached (ESD), and eclipsing detached (ED) systems with a false positive rate of only 4\%, 4\%, and 8\%, while complete to 64\%, 46\%, and 32\%, respectively. When classification confidence is relaxed, the EBF identifies and classifies ECs, ESDs, and EDs with a slightly higher false positive rate of 6\%, 16\%, and 8\%, while much more complete to 86\%, 74\%, and 62\%, respectively. Through our processing of the entire Kepler “Q3” data set, we also identify 68 new candidate EBs that may have been missed by the human generated Kepler EB Catalog. We discuss the EBF’s potential application to light curve classification for periodic variable stars more generally for current and upcoming surveys like K2 and the Transiting Exoplanet Survey Satellite.

Key words: binaries: eclipsing – methods: data analysis – methods: statistical – stars: variables: general

Online-only material: color figures

1. INTRODUCTION

Periodic variable stars, in particular eclipsing binary (EB) star systems, continue to be of central importance to a variety of applications in stellar astrophysics. EBs are crucial astrophysical benchmarks for stellar evolution theory as they can provide fundamental physical parameters of stars with very high precision and accuracy. Fundamental issues, from the stellar mass–radius relation using EBs in nearby clusters and the field (e.g., Lacy 1977), to the extragalactic distance scale using EBs in the Large Magellanic (Pietrzyński et al. 2013) and Small Magellanic Cloud (Graczyk et al. 2014), rely on the discovery and analysis of statistically significant samples of EBs and/or of rare but astrophysically important EBs. Indeed, EBs in which at least one component is of very low mass, in a scarce evolutionary phase, or itself a periodic variable are astrophysically very interesting and useful, yet rare and thus only small numbers of these benchmark systems are found in current catalogs. For example, there is only one known EB system comprising two brown dwarfs (Stassun et al. 2006, 2007), and only a couple of EBs containing a Classical Cepheid pulsator are currently known (e.g., Pietrzyński et al. 2010). Repositories of large time-series photometric survey data, such as the Kepler archive, provide an ideal sample from which to identify periodic variables including these interesting EBs for focused followup investigation. However, constraints on the rate of human analysis make mining this data using classical techniques prohibitive.

An intelligent data pipeline (IDP) that contains an automated light curve classifier to replace the human bottleneck is then necessary for the future pace of large survey data rates. Ideally such an IDP would be capable of automatically finding EBs in light curve data sets and then accurately classifying these EBs by morphological type (e.g., eclipsing contact (EC), eclipsing semi-detached (ESD), eclipsing detached (ED)) to enable focused followup analyses. Several algorithms have been developed for this purpose (e.g., Wyzykowski et al. 2003; Devor 2008; Graczyk et al. 2011; Devlinney et al. 2012 these are discussed further in Section 4), yet a fully automated pipeline remains an open task.

An automated classification pipeline with parameters adaptable to multiple time series photometric surveys would be immediately applicable to the Kepler data set (Christiansen et al. 2013) as well as the follow-up K2 mission (Howell et al. 2014) and would be well suited to the upcoming Transiting Exoplanet Survey Satellite (TESS, Ricker et al. 2014). In addition, the development of the Large Synoptic Survey Telescope (LSST) contains a requirement that data processing enable a fast and efficient response to transient sources (i.e., automated identification of variable stars and astrophysically interesting binaries) with a robust and accurate preliminary classification (Ivezić et al. 2011). In short, there is a large need currently and into the near future for methods that can reliably process large quantities of light curve data and automatically identify astrophysical objects of interest such as EBs with minimal human intervention.
So motivated, especially the immediate applicability of an automated pipeline for the study of EBs in K2 (Welsh et al. 2013), in this paper we present a fully automated, adaptive, end-to-end computational pipeline—the Eclipsing Binary Factory (EBF). The EBF first identifies EBs, and then morphologically classifies them as either EC, ESD, or ED. The classification module of the EBF pipeline was described in the first paper in this series (Paegert et al. 2014); here we add the remaining computational modules for a complete automated pipeline. In this paper we also then measure the accuracy and completeness of our EBF generated catalog of Kepler EBs against the manually constructed Kepler EB Catalog (Prša et al. 2011; Slawson et al. 2011).

The EBF takes a modular approach to automatically process large volumes of survey data into patterns suitable for recognition by an integrated, fully trained artificial neural network (ANN). This modular pipeline is designed to accept as input the direct archival data from a time-series photometric survey where each module’s parameters are adaptable to the specific characteristics of the data (e.g., photometric precision, data collection cadence, flux measurement uncertainty) and define the output options to produce the ANN generated catalog of EB classifications (i.e., EC, ESD, or ED). The computational efficiency of the ANN classifier, $O(2^N)$, where $N$ is the number of neurons in the neural network, is quite fast —of order 0.8 seconds for classification of ~50,000 light curves using a standard workstation. This then alleviates the current bottleneck by limiting follow-up human analysis to just the light curves automatically classified as EBs. In addition, the portability of this modular software pipeline makes the EBF a solution for the automated probabilistic identification and morphological classification of EBs in other large repositories of time-series, photometric data currently underway or planned for the near future.

The Kepler mission monitored ~200,000 targets brighter than the magnitude limit $Kp \sim 16$ in the Kepler spacecraft’s field of view (FoV). Kepler captured and summed these images as long cadence, 29.4 minutes, data files (Jenkins et al. 2010) that contain one quarter, or ~90 days, of observations across the FoV for sixteen quarters until reaction wheel failure placed the instrument into safe mode. The complete scientific data set, archived in MAST at the Space Telescope Science Institute, provides simple aperture photometry (SAP) with the photometric precision necessary to detect planetary transits (Koch et al. 2010) and astero-seismological signals (Gilliland et al. 2010) from solar-like stars: about 20 ppm for 12th magnitude G2V stars for a 6.5 hour integration (van Cleve & Caldwell 2009). As such, this data set provides both the appropriate sample size and precision for use in testing the EBF pipeline. In addition, the Kepler Eclipsing Binary Catalog Version 3 (KEBC3) of ~2600 manually classified EBs and a corresponding catalog of manually identified false positives in the Kepler FoV (Prša et al. 2011; Slawson et al. 2011) provides a benchmark of the EBF’s classification accuracy (i.e., false positive rate) and completeness (i.e., false negative rate). Moreover, full processing of the Kepler data set with the EBF automated pipeline provides the opportunity to uncover some additional EBs that may have been missed by the human generated KEBC3.

We demonstrate the efficacy of the EBF by automatically identifying and morphologically classifying EBs in the ~165,000 Kepler Quarter 3 (“Q3”) SAP light curves. In Section 2 of this paper we describe the procedures and methods of each EBF module. Section 3 presents the results of the EBF applied to the Kepler “Q3” light curves. Section 3.1.1 benchmarks the pre-classification data processing modules against the metrics of the manually corrected and phased light curves in the KEBC3 while Section 3.1.2 focuses on the EBF pipeline’s false positives and false negatives as a measure of classification accuracy and completeness. Finally, in Section 4 we discuss the future expansion of the EBF to include the automated identification and classification of other periodic variable stars such as $\delta$ Cepheid fundamental mode (DCEP-FU) and first overtones (DCEP-FO), $\delta$ Scuti (DSCT), RR Lyrae $ab$ (RRAB) and $c$ (RRC), and Mira (MIRA) as well the EBF’s applicability to other large time-series, photometric surveys. A set of newly identified candidate EBs that may have been missed by the human generated Kepler EB Catalog is provided in Appendix A. We conclude with a brief summary in Section 5.

### 2. METHODS: THE ECLIPSING BINARY FACTORY PIPELINE

The EBF Pipeline, diagrammed in Figure 1, is a compilation of five fully automated, adaptable software modules that process light curves from large astronomical surveys in order to...
automatically identify EBs and classify the each EB’s morphology. In this section we describe the parameters with which systematics are removed, periodic variable stars are identified, phased light curves are represented as ANN recognizable patterns, and constrain how EB candidate classifications are tuned and validated as a function of the survey instrument’s photometric precision, the data set duration and cadence, the flux measurement uncertainty, and the desired level of classification confidence, respectively. The procedure of each EBF module is described below in turn, then we test the overall pipeline performance by the rate of correctly classified systems and the completeness of the EBF classified sample as compared to the KEBCC3 benchmark sample classified by human analysis. We use the Kepler “Q3” long cadence data files as described in the Kepler Data Release 4 Notes KSCI-199044-001 (van Cleve 2010) to tool and test the EBF pipeline.

2.1. Data Correction Module

Corrections to the raw photometry collected from large survey instruments are generally necessary to remove systematic artifacts from the data. Though most surveys include their own data correction processes, there is no guarantee that the method employed preserves the necessary signal from which the EBF will identify and classify EBs. In the case of the Kepler “Q3” light curves corrections to the SAP flux are necessary to remove systematics such as differential velocity aberration, thermal gradients across the spacecraft, and pointing variations (Kinemuchi et al. 2012). Of course, the Kepler archive data reduction pipeline removes systematics in the “Q3” SAP data to produce pre-search conditioned data (PCD) using the method of cotrending basis vectors; however, PCD may lack the astrophysical signatures of variables whose harmonic thresholds were not met (McQuillan et al. 2012). Hence, we employ a data correction module (DCM) to remove trending artifacts while preserving the necessary signal as well as to normalize the flux in order to present the remainder of the pipeline with a standardized, detrended light curve. The DCM step is applied to all of the input light curves (i.e., no light curves are filtered out at this stage).

The DCM accepts as input the time-series flux with the associated uncertainty of each data point and detrends the flux via the sigma-clipping algorithm of Slawson et al. (2011); where data points outside a standard deviation interval of a least squares Legendre polynomial fit to order $l$, expressed by the generating function

$$P_l(x) = \frac{1}{2^l l!} \frac{d^l}{dx^l} \left[ (x^2-1)^l \right],$$

are iteratively discarded until there are no remaining data points outside the interval. For each data point input, the DCM returns a point containing a normalized flux value with a new uncertainty. The EBF’s implementation of this method allows the parameters of the module to be set to the specific characteristics of a data set by defining the sigma clipping thresholds and Legendre polynomial order. Additionally, the module allows for the input of the time intervals of known breaks in the light curves due to instrument operations. In this paper we detrend the Kepler “Q3” SAP light curves using asymmetric sigma-clipping thresholds of $(-1\sigma, 3\sigma)$ with the 10th order Legendre polynomial

$$R_0(x) = 256^{-1}(46,189x^{10} - 109,395x^8$$

$$+ 90,090x^6 - 30,030x^4 - 3,465x^2 - 63)$$

and known breaks in the “Q3” light curves beginning at BJD 281.0, 291.0, and 322.5. Figure 2 shows the uncorrected Kepler “Q3” SAP light curve for a target and the EBF detrended, normalized light curve where systematics were removed by the DCM and the red circles indicate data gaps given by the aforementioned breaks.

2.2. Harmonics Recognition Module (HRM)

The EBF pipeline is currently optimized for identification and classification of EBs, which exhibit periodically varying light curves. Therefore the precursors to morphological classification of EBs via pattern recognition are the identification of a target’s periodic variability and a subsequent phase folded light curve. Hence, the detrended normalized light curve

Figure 2. Left: Kepler “Q3” SAP light curve for target KIC 1026032. Right: The EBF detrended, normalized light curve for target KIC 1026032 where systematics were removed by the data correction module using the sigma-clipping algorithm of Slawson et al. (2011) for asymmetric sigma-clipping thresholds $(-\sigma, 3\sigma)$ with a least squares fit to the 10th order Legendre polynomial. The red circles indicate data gaps that are accounted for by the pipeline.

(A color version of this figure is available in the online journal.)
is next processed by the HRM where periodic variability is probed by the analysis of variance (AoV) method (Schwarzenberg-Czerny 1989; Devor 2005) included in the VAR-TOOLS1.202 package (Hartman et al. 2008). The HRM takes as input all of the normalized light curves output by the preceding DCM, but outputs only those light curves that are deemed to exhibit truly periodic behavior (of any kind, i.e., the output light curves from the HRM stage include EBs but also any other type of periodic variable).

Two separate searches for the first three harmonics, each with a coarse sub-sample setting of 0.1 and a fine-tune setting of 0.01, are conducted over the entire data set. The first search separates the data into five bins and the second uses fifty bins such that skewed harmonics from bin size related flux variance is minimized. We then compare the strongest harmonic signal from both periodograms and, in the event of disagreement between the primary periods, assign the period (along with all other associated AoV statistics) from the periodogram with the smallest false alarm probability.

To eliminate light curves that are unlikely to exhibit truly significant periodic variability, we use the AoV calculated signal to noise ratio (S/N). Specifically, periodograms with a S/N ≤ 100 are rejected by the module while the remaining light curves, now identified as candidate periodic variables, are accepted. These accepted candidates’ light curves are phased by the standard equation

$$\rho = \frac{t - t_e}{p} \quad \text{int} \left( \frac{t - t_e}{p} \right) \quad (3)$$

where ρ is the phase of an associated time value, t, for a given reference time, t_e, and the assigned period, p.

For an EB, we want the reference or zero phase ρ_e, where t = t_e, to be associated with the primary eclipse. However, not all of the periodic variables identified by the HRM will be EBs; there will be other types of periodic variables where the most physically meaningful reference phase is not an eclipse but rather a flux maximum. Therefore, to maintain generality, we associate ρ_e with either the minimum or maximum flux, dependent upon whether the light curve spends most of its time above or below the median flux. Most true EBs will have flux points in the “high” state and will spend a minority of the time in eclipse. Here the determination is made by the value of the flux ratio (FR) as described by Coughlin et al. (2011):

$$FR = \frac{\text{Maximum Flux} - \text{Median Flux}}{\text{Maximum Flux} - \text{Minimum Flux}} \quad (4)$$

By definition, the FR has a value on the interval (0, 1). For reference, a perfect sinusoid will have FR = 0.5. We assign ρ_e to the flux minimum, indicative of an EB, for an FR less than a defined threshold, and a flux maximum assigned at ρ_e otherwise.

The HRM’s parameters that are adaptable to the data set are the period search interval, based on the duration of the survey data, and the FR threshold. The latter is based on the tolerance for ECs, which in general are not perfect sinusoids and whose FR may be greater than 0.5 if they spend more than half the time above the median flux. As we limited the Kepler input files to “Q3” (∼90 days), the harmonics search was limited to a period range of 0.11–20.1 days in order to ensure at least three primary eclipses would be present in each light curve. In addition, we set the FR threshold to 0.7 based on a simple analysis of KEBC3 FRs in order to include those ECs whose light curves are not perfectly sinusoidal. Figure 3 shows a plot of the FR distribution in the KEBC3 versus cataloged morphology, and a “Q3” target as an example of an imperfect sinusoidal phased EC light curve with a FR > 0.5. An example of a phased ESD light curve, and an example of a phased ED light curve, each with FR < 0.5, are shown in Figure 4.

2.3. Pattern Generation Module (PGM)

As described in Section 2.4, the EBF’s classification of the phased light curves for the candidate periodic variables output from the HRM is dependent upon the ANN evaluating these light curves as linearly separable patterns. In its simplest form linearly separable means two sets of points in a plane may be divided by at least one line such that each data point belonging to one set is on one side of the line and each data point belonging to the second set is on the other side of that line.

Figure 3. Left: flux ratio dispersion in the KEBC3 vs. the cataloged morphology, where the cyan dots are EDs, the yellow dots are ESDs, and the blue dots are ECs. The red line marks the KEBC3 threshold for imperfect sinusoidal light curves. Right: EBF phased, detrended light curve of KIC 9004380 as an example of an imperfect sinusoidal EC light curve where 0.5 < FR < 0.7.

(A color version of this figure is available in the online journal.)
When generalized to an $n$-dimensional space, linear separability is formalized such that every point $x \in \mathbf{X}_0$ satisfies $\sum_{i=1}^{n} w_i x_i > k$ and every point $x \in \mathbf{X}_1$ satisfies $\sum_{i=1}^{n} w_i x_i < k$, where $\mathbf{X}_0$ and $\mathbf{X}_1$ are two sets of points in an $n$-dimensional Euclidean space and $x_i$ is the $i$th component of $\mathbf{X}_j$ for $n + 1$ real numbers $w_1, w_2, ..., w_n, k$. The PGM therefore represents the phased light curves via piecewise smooth polynomial chains $\mathbf{X}_j$ to pass to the ANN as linearly separable patterns.

Since EBF efficiency goes as $O(N^2)$, where $N$ is the number of elements in $\mathbf{X}_j$, it becomes computationally expensive to process light curves possessing on the order of $10^3$ data points. Therefore, we condense the light curve by phase binning down to the order $10^2$ data points over the phase interval $(-0.5, 0.5)$, with the primary eclipse centered at $\rho_{\text{cen}}$. In addition to meeting the efficiency requirements, the phase binning process also minimizes $\rho_{\text{err}}$ error by smoothing over uncorrected systematics that may present a stray minimum flux below the true primary eclipse depth.

The PGM therefore fits the phase binned curves to piecewise smooth second order polynomial chains using the polyfit algorithm of Prša et al. (2008). However, the ANN’s use of the sigmoid activation function (see Section 2.4) leads to errors in processing large polynomial coefficients that may arise during polynomial fitting. Hence, we modify the polyfit algorithm to fit polynomial chains $\mathbf{X}_j$ with components $x_i$ described by three points in a two-dimensional Euclidean space (phase, flux). These three data points represent the chain’s two end-points, or knots, and the midpoint. Here each chain shares one knot with the next, thus a single chain is effectively described by two points in a plane. We add the additional modification of allowing either a four-polynomial chain $\{\mathbf{X}_0, \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3\}$ or two-polynomial chain $\{\mathbf{X}_0, \mathbf{X}_1\}$ fit based on a $\chi^2$ goodness of fit criterion. With these modifications the PGM processes phased light curves initially containing on order $10^3 \sim 10^4$ data points into at most four linearly separable patterns $\mathbf{X}_j$ where $N = 4$ for each $j = 1, 2, 3, 4$. That is, we reduce the recognizable pattern of the phase folded light curve which may contain a very large number of data points to a maximum of sixteen points.

Bin size selection is an adjustable parameter. This option allows for a condensed light curve based on the survey cadence and/or duration in order to manage the point density per bin. Of course, more bins lead to longer run times for the PGM and fewer bins lead to sparse light curves where the narrow eclipses of wide EBs might be smoothed over or missed. The minimum number of points per bin may also be adjusted in an attempt to capture narrow eclipses; however, from trial and error we find that no fewer than 50 phase bins and no fewer than three bins per chain is optimal.

To fit the Kepler “Q3” phased light curves we bin each light curve’s 5,000 long cadence data points into 200 (size = 0.005) phase bins with a polynomial chain that contains no fewer than three data points. The EBF generated binned light curves and linearly separable patterns, plotted as examples of Kepler EBs with the morphology EC, ESD, and ED are shown in Figures 5–7, respectively.

### 2.4. Artificial Neural Network Classifier (Pattern Recognition Module)

The EBF’s trained, validated, and tested ANN is the single hidden layer feed forward perceptron network diagrammed in Figure 8. This module takes as input the PGM output set of data points as a simple representation of the target light curve. These patterned inputs are recognized by the ANN and classified into the EB categories that it has been trained to recognize. The outputs of the ANN are therefore the classification of each light curve with an associated confidence level in the classification. A detailed description of the ANN, including its training, validation, and output classifications, is described in Paegert et al. (2014). Here we provide a summary of its specifications and functions.

The ANN functions as a two-stage pattern recognition classifier based on nonlinear statistical models. The input layer, nodes $x_1$ to $x_n$, are populated with the PGM fitted polynomial chains $\mathbf{X}_j$ as an input vector of $n$ components that is then propagated via weighted connections $w(h)$ to the hidden layer.
The activation function \( a(h) \) is in the form of the sigmoid
\[
\sigma(\nu) = (1 + e^{-\nu})^{-1}
\] (5)
where \( \nu \) is the variable used by the ridge function to derive the features \( H_1 \) to \( H_m \) as linear combinations of the inputs. The signals are again weighted by \( w(o) \) and propagated to the output layer where there are \( o \) target classifications coded as a 0 to 1 categorical target variable \( \nu \) for the \( o \)-th classification using the same sigmoid activation function to model the target nodes \( y_1 - y_o \) as linear combinations of \( H_1 \) to \( H_m \), where \( n = 19, \ m = 20, \) and \( o = 10 \). In other words, the ANN accepts \( n = 19 \) input parameters—representing the 16 polynomial chain parameters from the PGM as well as the logarithm of the period in days, the total \( \chi^2 \) as a measure of fit, and the maximum amplitude of the light curve’s normalized flux—and outputs a probability associated with each of the \( o = 10 \) known light curve classifications. The 10 known classifications are the three EB categories (EC, ESD, ED), the other variable types (DCEP-FU, DCEP-FO, DSCT, RRAB, RRC, MIRA), and a miscellaneous classification (MISC).

Training of the ANN was conducted by back-propagation, and gradient descent was used to minimize the error in the set of all weights \( \theta \), or \( err_\theta \), where the error function is given by
\[
err_\theta = R(x, y) = \sum_{i=1}^{n} \sum_{j=1}^{o} (y_{ij} - f(x_{ij}))^2
\] (6)
for \( i = 1, 2, 3, \ldots n \) and \( j = 1, 2, 3, \ldots o \). \( f(x_{ij}) \) is described by the softmax function
\[
f(x_{ij}) = g(T_k) = \frac{e^{T_k}}{\sum_{k=1}^{o} e^{T_k}}
\] (7)
as \( T \to \infty \), and was used to approximate the nonlinear input functions. As described in Paegert et al. (2014), training and validation of the ANN was done using 32,278 light curves from the Automated All-Sky Survey (ASAS; Pojmanski 1997).
These tests showed the ANN is capable of an accuracy of 90.6%, 76.5%, and 90.3%, and a completeness of 95.8%, 63.7%, and 87.8% for ECs, ESDs, and EDs, respectively.

In summary, for each patterned light curve from the PGM, the ANN outputs a most likely classification and an associated confidence in that classification. The output classifications for EBs can be either EC, ESD, or ED. Because the ANN was originally developed on the ASAS sample (Paegert et al. 2014), there is also a classification of MISC but secondary classification of EC, ESD, or ED.

2.5. Candidate Validation Module (CVM)

Once ANN classifications are complete, candidate EC, ESD, and ED light curves are filtered for false positives by the CVM. We position the CVM as a post classification process to provide the option to examine some or all of the filtered candidates.

The CVM filters candidates by two distinct metrics. The first is the polyfit algorithm’s total goodness of fit, $\chi^2_T$, such that

$$\chi^2_T \equiv \sum_{j=0}^{n} \sum_{i=1}^{m} w_{ji} (x_{ji} - \Phi_{ji})^2;$$

where $\Phi_{ji}$ is the normalized, detrended flux of phase bin $l$ and $x_{ji}$ is the $i$th component of the $j$th polynomial chain $X_j$ for $j = 0, 1, \ldots, m$ that satisfies the requirement of linear separability from Section 2.3. The $\chi^2_T$ metric describes how well the PGM represents the candidate light curves and is a useful measure of whether the ANN’s recognition of the EBF generated pattern translates to an accurate classification as an EB.

The second metric is the normalized, detrended light curve’s scaled sum of all second derivatives for the portion of the curve between each and every data point, which we refer to as $\delta^2$ and defined as

$$\delta^2 \equiv \frac{1}{\Phi_{\max} - \Phi_{\min}} \times \sum_{i=2}^{n-1} \frac{1}{\rho_{i+1} - \rho_{i-1}} \left( \frac{\Phi_i - \Phi_{i-1}}{\rho_i - \rho_{i-1}} - \frac{\Phi_{i+1} - \Phi_i}{\rho_{i+1} - \rho_i} \right)$$

where $\rho_i$ and $\Phi_i$ are the phase and flux for the $i$th of $n$ data points. As such, $\delta^2$ is a measure of the dispersion in the detrended light curve and when compared to the average detrended flux measurement uncertainty calculated in the DCM provides a criterion for filtering a pseudo-periodic signal that may arise from random noise.

The CVM adapts the $\chi^2_T$ and $\delta^2$ filter metrics by using ~30 known EBs contained in the survey. The module then calculates the upper limits on both the $\chi^2_T$ and $\delta^2$ from this benchmark subset. A CVM analysis of the KEBC3 EBs contained in the Kepler “Q3” data set and recovered by the ANN provides a $\chi^2_T$ upper limit of 100 and a $\delta^2$ upper limit of 1200 as shown by the histograms in Figure 9.
Figure 10 shows the light curves of two example Kepler targets where the PGM fits to some quasi-systematic artifact that the ANN consequently mis-classifies as an ESD, however the CVM criteria of $\chi^2_T \geq 100$ filters this candidate as a false positive. Right: EBF generated light curve of KIC 6774862 where the PGM fits to a pseudo-sinusoidal signal that the ANN appropriately classifies as an EC, however the CVM filter criteria of $\delta^2 \geq 1200$ prevents this false positive from populating the automated EB catalog. (A color version of this figure is available in the online journal.)

The Astronomical Journal, 148:125 (14pp), 2014

Parvizi, Paegert, & Stassun

3. RESULTS

We use the 2612 EBs cataloged in the KEBC3 to benchmark the EBF pipeline outputs as an initial assessment of pipeline capabilities on standard morphology ECs, ESDs, and EDs. Thus, we first remove from the benchmark sample those KEBC3 EBs flagged as eclipse timing variations (ETV) induced by third body candidates (Conroy et al. 2013), heart beat (HB), and those classified in the KEBC3 as uncertain (UNC). We also exclude those KEBC3 EBs with cataloged periods outside the Kepler “Q3” duration parameters of $0.11 \text{ days} \leq \text{period} \leq 20.1 \text{ days}$ and of course those cataloged EBs that do not appear in the Kepler “Q3” data set. Thus, we benchmark the EBF pipeline against an adjusted KEBC3 sample of 1198 cataloged EBs.

Figure 10. Left: histogram of the PGM $\chi^2_T$ over the ANN recovered sample of KEBC3 EBs used to calculate the CVM filter criteria of $\chi^2_T \geq 100$. Right: histogram of the CVM $\delta^2$ over the ANN recovered sample of KEBC3 EBs used to calculate the CVM filter criteria of $\delta^2 \geq 1200$.
(A color version of this figure is available in the online journal.)
3.1. Eclipsing Binary Factory Performance Metrics

3.1.1. Pre-classification Metrics from the Data Correction Module, Harmonics Recognition Module, and Pattern Generation Module

We compare the metrics of the EBF pipeline’s pre-classification output to the same metrics of the manually determined KEBC3. Note that the KEBC3 light curves contain ~50,000 data points from more than 10 quarters of observations while the “Q3” observations include only ~5000.

We evaluate the DCM’s performance by comparing the 200 phase binned, normalized flux values of each common target. We find the DCM’s performance to be in excellent agreement, mean $\chi^2 = 0.0012$, with the KEBC3 benchmark. This DCM light curve agreement breaks down to a mean $\chi^2$ of 0.00373, 0.00264, and 0.00052 for cataloged ECs, ESDs, and EDs respectively.

The HRM assigned period agrees to within ±0.09 days for 99.67% of the EBF identified periodic variable candidates found in the KEBC3, though the HRM failed to identify statistically significant periodicity in 5.59% of the benchmark sample, and therefore were not considered for follow-up classification by the ANN. A further comparison of the HRM assigned periods to those manually determined in the KEBC3 reveals a period alias rate of 21.65% (i.e., the percentage of EBF determined periods that are an alias of the KEBC3 period). However, aliasing does not negatively impact ANN pattern recognition and the true period is easily determined by examining the candidate light curve. Figure 11 shows examples of both the EBF generated and the manually determined KEBC3 light curves for targets where the EBF assigns a half-period alias. It was also discovered that phase disagreement was common for targets with FR > 0.5. Though there is no adverse effect on pattern recognition for the example targeted EC in Figure 12, the ambiguity in phasing a light curve with an FR of ≤0.7 on the minimum flux led to diminished pipeline accuracy as other variable types mistaken by the pipeline for EBs (i.e., δ-cepheids incorrectly phased on the flux minimum) contributed to the false positive rate.

The PGM fitted the EBF phased, detrended light curves with either four- or two-polynomial chains and a goodness of fit described by a median $\chi^2$ of 5.39 as compared to the unaltered polyfit algorithm used to fit the KEBC3 samples with a median $\chi^2$ of 3.74. This goodness of fit includes cases where the PGM patterned fit is in excellent agreement with the manually determined KEBC3 light curves, such as the target shown in Figure 13, as well as instances where the patterned fit fails to capture the full depth of the eclipses as circled in Figure 14.

The subtle deviation between a phased light curve and an EBF generated pattern may present a challenge to an automated solution estimator, yet the PGM’s representation of the light curves shape allows the ANN to successfully recover and classify these systems. This and other deviations, such as the small aberrant third chain circled in red on Figure 11, are the result of the three data points per chain minimum parameter we adopted in the PGM in order to capture narrow eclipses. Though ~10% of the non-recovered KEBC3 benchmark sample is comprised of wide EDs with narrow eclipses that contain less than three binned data points, the three data point rule remains a hard minimum to ensure pattern generation integrity. It is important to note that the PGM failed to pattern 3.33% of the benchmark sample and that these light curves were not considered for follow-up recognition by the ANN.

3.1.2. Classification Metrics from the Artificial Neural Network and Candidate Validation Module

The performance of the ANN for cataloged binaries in the Kepler “Q3” data set is benchmarked by KEBC3 EBs with period $p$ where 0.11 days ≤ $p$ ≤ 20.10 days while ETVs, HBs, and UNCs were excluded. The ANN reports each of its classifications with a posterior probability $P$ interpreted as a level of confidence for the morphological class of EC, ESD, ED, or any other periodic variable class contained in the ASAS training set (Paegert et al. 2014).

Before discussing the performance of the EBF pipeline as a whole, we consider the ANN performance alone. This is evaluated only for those light curves passed forward by the DCM, HRM, and PGM such that the ANN performance metrics are not hindered by the exclusion of light curves by the
preceding modules. The ANN’s classification accuracy metric is calculated using the KEBC3 targets flagged in the KEBC3 database as InCat:False (ICF), a false positive EB (e.g., blends, planets, pulsators, rotators). Thus we only select those targets designated ICF that lack a cataloged morphology (e.g., pulsators, rotators) as an inaccurate classification. That is, the ANN’s accuracy is given as the percentage of ANN EC, ESD, ED classifications that are Kepler targets currently cataloged in the KEBC3 as an EB as opposed to those currently cataloged as a pulsating or rotating variable. The ANN’s completeness metric is simply the percentage of the adjusted KEBC3 benchmark sample that the ANN successfully recovered as an EB. Using the Kepler “Q3” targets, the ANN accuracy is found to be 93.68%, 84.27%, and 83.03% with a completeness of 94.17%, 93.62%, and 93.63% for EC, ESD, and ED, respectively. The CVM’s performance was evaluated by a manual inspection of the 91 high-confidence ($P \geq 0.90$) EB candidate light curves generated by the EBF. Here we remove from these candidates 21 Kepler “Q3” targets that suffer from the phasing ambiguity discussed in Section 3.1.1. This manual inspection of the remaining 68 candidates found that the high-confidence CVM validated EBs are likely reliable to ~90% accuracy, commensurate with the integrated EBF pipeline’s accuracy as discussed below.

3.1.3. Overall Performance

To evaluate the performance of the integrated EBF pipeline, we describe the EBF generated catalog for candidate ECs, ESDs, and EDs with two distinct queries in order to evaluate EBF accuracy and completeness in the Kepler FoV both at high-confidence ($P \geq 0.90$) and with no regard to confidence ($P \geq 0.01$). Here, no regard to confidence means either the primary ANN classification is EC, ESD, ED, or the primary classification is the placeholder MISC with a secondary

![Figure 12](image1.png)

Figure 12. Left: KEBC3 light curve vs. EBF phased, binned light curve of KIC 9004380 detrended by sigma-clipping over (-1σ, 3σ) interval of the Legendre function $P_0(x)$ with 200 phase bins. Right: KEBC3 light curve vs. EBF patterned curve of KIC 9004380 fit to polynomial chains of no less than three bins. Here the EBF recovers the cataloged EC in spite of a $\rho_0$ disagreement with the manually adjusted benchmark.

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

![Figure 13](image2.png)

Figure 13. Left: KEBC3 light curve vs. EBF phased, binned light curve of KIC 7258889 detrended by sigma-clipping over (-1σ, 3σ) interval of the Legendre function $P_0(x)$ with 200 phase bins. Right: KEBC3 light curve vs. EBF patterned curve of KIC 9004380 fit to polynomial chains of no less than three bins. Here the EBF recovers the cataloged ESD with excellent agreement between the KEBC3 light curve and the EBF generated pattern.

(A color version of this figure is available in the online journal.)
classification of EC, ESD, or ED with any posterior probability on the interval \((0, 1)\). The accuracy and completeness for each condition are listed in Table 1.

These two queries serve to exemplify the range of utility in the EBF’s automated output of candidate Kepler FoVs EBs. For example, a high-confidence \((P \geq 0.90)\) query for ECs in the interest of analyzing the characteristics of a homogeneous sample of Kepler ECs, with a high certainty of morphological classification, recovers \(\sim 64\%\) of those available in the “Q3” data set with only \(\sim 4\%\) false positives. Conversely, a query with no regard to confidence \((P \geq 0.01)\) in the interest of a statistical analysis using a large number of Kepler ECs, without regard to the certainty of their morphology \((P \geq 0.01)\), recovers \(\sim 86\%\) of those in the “Q3” data with \(\sim 6\%\) false positives.

### 3.2. New High Confidence Candidate Eclipsing Binaries Identified with the Eclipsing Binary Factory

In addition, for the high-confidence condition we find 68 CVM validated EB candidates that do not appear in the KEBC3 catalog. The set of these newly identified candidates are provided in Appendix A. In general, these newly identified candidates are in many cases visually more complex and less obviously identified as EBs via manual visual inspection of the light curves. This is not surprising, as the EBs previously cataloged in the KEBC3 were by definition those EBs that could be readily identified by visual human analysis, and therefore these new candidates are only those that would have been most likely to be missed by human visual inspection. Nonetheless, visual inspection of these new candidates identified in an automated manner by the EBF indicates that most if not all of these are bona fide EBs. In any case, these new candidates will be examined further in a follow-up paper.

### 4. DISCUSSION

Previous works have developed aspects of an automated EB classifier, although to date a fully automated end-to-end pipeline has remained an open challenge. For example, the EB identification algorithm of Wyrzykowski et al. (2003) automatically identified candidates in the Optical Gravitational Lensing Experiment (OGLE), but then relied on visual inspection of all candidate light curves for manual classification. Similarly, the semi-automated method of Graczyk et al. (2011) relied on visual inspection to manually remove artifacts from the OGLE-III candidate light curves. Though the pipeline of Devor (2008) automatically filtered 97% of the Trans-Atlantic Exoplanet Survey (TrES) data as non-periodic, a manual removal of 86% of the remaining light curves was necessary to finalize the catalog. More advanced algorithms described by Devinney et al. (2012), such as the Eclipsing Binaries via Artificial Intelligence (EBAI) method (Prša et al. 2008), used sophisticated ANNs to estimate the characteristic parameters of detached EBs in the Kepler Field; however, the classification of EB morphological sub-classes remained largely a manual task.

In contrast to the algorithm of Wyrzykowski et al. (2003), where the authors are only able to estimate that their EB catalog is complete to \(\sim 85\%\) with an efficiency of \(\sim 95\%\), and where the classification of EB sub-classes is achieved only through visual inspection, the EBF’s completeness and accuracy rates are precisely benchmarked by the KEBC3 and based upon the ANN’s fully automated classification of sub-classes.

---

**Table 1**

| Query | EC | ESD | ED |
|-------|----|-----|----|
| \(P \geq 0.90\) for \(0.11 \text{ days} \leq \text{period} \leq 20.10 \text{ days}\) | 96.12 | 95.53 | 91.93 |
| \(P \geq 0.01\) for \(0.11 \text{ days} \leq \text{period} \leq 20.10 \text{ days}\) | 93.68 | 84.27 | 83.03 |

**Note.**

*Classifications with no regard to confidence. Either the primary ANN classification is EC, ESD, ED, or the placeholder MISC with a secondary classification of EC, ESD, or ED with any posterior probability on the interval \((0, 1)\).*
Additionally, the EBF’s single-layer ANN represents a much simpler representation of the classifier algorithm with a more robust training set (i.e., 19 input nodes and trained on ∼32,000 distinct ASAS exemplars). The Wyrzykowski et al. (2003) method requires the conversion of OGLE photometric light curves into representational pixel images that are then analyzed by a 1050 input node three-layer ANN trained on variations of only 10 OGLE EB exemplars.

The OGLE-III analysis of Graczyk et al. (2011) is based on a 13 step algorithm that, when complete, reports a false-positive rate of ∼10% and thus requires a manual inspection of each and every EB candidate light curve. In contrast, the EBF’s automated filter reduces false positives to less than ∼10%. The semi-automatic method of Devor (2008) also requires manual classification of the EB sub-classes and is further constrained to the analysis of ED and ESD TrES light curves while the EBF extends the result to include the EC class. The EBF pipeline again represents an advance in full automation and adds the ability to automatically classify the EC sub-class when compared to more advanced solution estimators, such as the EBAI method (Prša et al. 2008).

The purpose of the EBF pipeline is to alleviate the human bottleneck in discovering astrophysically interesting EB systems by automatically identifying and classifying these systems from large volumes of survey data as a foundation for an IDP designed for the future of astroinformatics. The performance results using the Kepler “Q3” Data Release demonstrate the EBF’s accuracy for those EBs flagged at high confidence, however the completeness metric leaves room for improvement. An examination of the pipeline’s unclassified KEBC3 EBs reflects the 5.59% exclusion rate of the AoV-S/N output filter, as well as the 3.33% failure rate of the PGM. Here we find that 8.92% of the benchmark sample never reaches the ANN. In addition, we discover that a simple examination of the PR in relation to an arbitrary threshold of ∼0.5 is insufficient to discriminate between ECs and other variable types. Thus, the necessity for the pipeline to first reliably identify harmonics as well as correctly phase and fit the input data without manual adjustment is obviously an area in need of refinement.

These period and subsequent phasing errors can potentially be mitigated in the future by layering the HRM with multiple harmonic search methods for varied period length. For example, the computationally inexpensive Fast Chi Squared (Fχ2) method of Palmer (2009) and the Boxed Least Squares transit search algorithm (Kovács et al. 2002) may supplement the AoV method. This multi-layered harmonics search could be combined with a revised PGM where a post-fitting χ2 analysis may be used to detect errors and validate assigned periods. Phasing may then be subjected to other constraints (e.g., the symmetry versus...
asymmetry of the primary eclipse ingress and egress for those targets with FR ~0.5. Training deficits such as the lack of a sufficient number of wide EB light curves with narrow transits in the ASAS exemplars as well as an inherent morphological ambiguity in binary systems can potentially be rectified with further training, testing, and validation using expanded survey data sets.

The EBF’s adaptability, facilitated by the modular parameters described in Section 2, and the automated catalog’s query options, demonstrates the pipeline’s utility for a variety of applications. Applying the EBF to other current time-domain photometric surveys, such as K2 and the upcoming TESS mission should provide similar results. In addition, the EBF’s ANN is currently tooled and trained to recognize and classify the other types of periodic variables discussed in Paegert et al. (2014; i.e., the ASAS periodic variable classifications DCEP-FU, DCEP-FO, DSCT, RRAB, RRC, and MIRA). Only a benchmark of variables in the Kepler FoV, parallel to the KEBC3, is needed to tool and test the EBF’s expansion into automatically cataloging these and other types of periodic variables.

5. CONCLUSION

The EBF generated a fully automated catalog of EB systems in the Kepler FoV from the Kepler “Q3” Data Release. This catalog, benchmarked by the manually generated KEBC3, identified and classified to 90% certainty EC, ESD, and ED systems (0.11 days ≤ period ≤ 20.10 days) with an accuracy of 96.12%, 95.53%, and 91.93%, while complete to 63.93%, 45.80%, and 32.40% respectively. When classification
certainty is not considered, the EBF catalog identified and classified ECs, ESDs, and EDs (0.11 days $\leq \text{period} \leq 20.10$ days) with an accuracy of 93.68%, 84.27%, and 91.93%, while complete to 86.07%, 73.95%, and 61.59% respectively. In addition, the EBF identified 68 new candidate EBs that may have been missed by the human generated KEBC3 (see Appendix A).

This demonstrates the EBF pipeline as an effective alternative to the constraints of human analysis when probing large time-domain, photometric data sets. As such, the EBF computational pipeline is a viable framework for the development of a more complex IDP containing a follow-up automated solution estimator that would enable the automatic determination of physical EB parameters. Future work on the EBF will include enhancing the harmonics identification capabilities with multiple search methods and phasing constraints beyond the FR, a revised pattern generator to constrain polynomial chain fits to smaller errors, and more robust training for the ANN on wide binary systems. In addition, an expansion of EBF searches in the Kepler data, and in the upcoming K2 data, to include the periodic variables of type $\delta$-cepheid (fundamental and first overtones), $\delta$-scuti, RR-Lyrae ($ab$ and $c$), and Mira will be published in follow-up papers. Over the long term, the EBF general classification approach can be adopted by future surveys such as TESS and LSST.

We are grateful to A. Prša, K. Conroy, and all those at Villanova University who contribute to the Kepler Eclipsing Binary Catalog, to J. Pepper for his insights and experience with large surveys, and to R. Siverd for his expert assistance with harmonics. Kepler was selected as the 10th mission of the Discovery Program. Funding for this mission is provided by NASA, Science Mission Directorate.

APPENDIX

ECLIPSING BINARY FACTORY HIGH-CONFIDENCE NEW ECLIPSING BINARY CANDIDATES

In Table A1, we provide the KIC, R.A., decl., period, and Kepler magnitude for 68 high-confidence ($P \geq 0.9$), astrophysically interesting EB candidates newly identified by the EBF pipeline from the Kepler “Q3” light curves that are not currently included in the human generated KEBC3. Figures A.1–A.3 represent a sample of “Q3” candidate light curves. The list of candidates also includes three targets with a current disposition of “CANDIDATE” in the Kepler Objects of Interest (KOI). Figure A4 shows the light curves for KOI K00188.01, K00830.01, and K00183.01.

REFERENCES

Christiansen, J. L., Jenkins, J. M., Caldwell, D. A., et al. 2013, in Kepler Data Characteristics Handbook (KSCI-19040-004) (CA: NASA Ames Research Center, Moffet Field), http://archive.stsci.edu/kepler/manuals
Conroy, K., Prša, A., Stassun, K. G., et al. 2013, AJ, 147, 45
Coughlin, J., López-Morales, M., Harrison, T. E., Ule, N., & Hoffman, D. I. 2011, AJ, 141, 78
Deviney, E. J., Prša, A., Guinan, E. F., et al. 2012, in Proc. IAU Symp. 282, From Interacting Binaries to Exoplanets: Essential Modeling Tools, ed. M. T. Richards & I. Hubeny (Cambridge: Cambridge University Press), 321
Devor, J. 2005, ApJ, 628, 411
Devor, J. 2008, AJ, 135, 850
Gilliland, R. L., Brown, T. M., Christensen-Dalsgaard, J., et al. 2010, PASP, 122, 131
Graczyk, D., Pietrzynski, G., Thompson, I. B., et al. 2014, ApJ, 780, 59
Graczyk, D., Sozyński, R., Poloński, G., et al. 2011, AcA, 61, 103
Hartman, J. D., Gaudi, B. S., Holman, M. J., et al. 2008, ApJ, 675, 1233
Howell, S. B., Sobeck, C., Haas, M., et al. 2014, PASP, 126, 398
Ivezić, Z., Tyson, J. A., Acosta, E., et al. 2011, arXiv:0805.2366
Jenkins, J., Caldwell, D. A., Chandrasekaran, H., et al. 2010, ApJL, 713, L160
Kimemuchii, K., Fanelli, M., Pepper, J., et al. 2012, PASP, 124, 919
Koch, D., Borucki, W. J., Basri, G., et al. 2010, ApJL, 713, L79
Kovács, G., Zucker, S., & Mazeh, T. 2002, A&A, 391, 396
Lacy, C. H. 1977, ApJ, 218, 444
McQuillan, A., Aigrain, S., & Roberts, S. 2012, A&A, 539, A137
Paczynski, B., Szczygieł, D. M., Pilecki, B., & Pojmanski, G. 2006, MNRRAS, 368, 1311
Paegert, M., Stassun, K. G., & Burger, D. M. 2014, AJ, 148, 31
Palmer, D. 2009, ApJ, 695, 496
Pepper, J., Stanek, K. Z., Pogge, R. W., et al. 2008, AJ, 135, 907
Pietrzyński, G., Graczyk, D., Gieren, D., et al. 2013, Natur, 492, 76
Pietrzyński, G., Thompson, I. B., et al. 2010, Natur, 468, 542
Pojmanski, G. 1997, AcA, 47, 467
Prša, A., Batalha, N., Lawson, R. W., et al. 2011, AJ, 141, 83
Prša, A., Guinan, E. F., Deviney, E. J., et al. 2008, ApJ, 687, 542
Ricker, G. R., Winn, J. N., Vanderspek, R., et al. 2014, BAAS, 42, 459
Schwarzenberg-Czerny, A. 1989, MNRRAS, 241, 153
Slawson, R., Prša, A., Welsh, W. F., et al. 2011, AJ, 142, 160
Stassun, K. G., Mathieu, R. D., & Valenti, J. A. 2006, Natur, 440, 311
Stassun, K. G., Mathieu, R. D., & Valenti, J. A. 2007, ApJ, 664, 1154
van Cleve, J., & Caldwell, D. 2009, Kepler Reprint Manual (KSCI-10933-001) (CA: NASA Ames Research Center, Moffet Field), http://archive.stsci.edu/kepler/manuals
van Cleve, J. 2010, in Kepler Data Release 4 Notes (KSCI-10944-001) (CA: NASA Ames Research Center, Moffet Field), http://archive.stsci.edu/kepler/manuals
Welsh, W., Bloomer, S., Conroy, K., et al. 2013, arXiv:1309.1176
Wyrzykowski, L., Udalski, A., Kubiak, M., et al. 2003, AcA, 53, 1