

Identifying Exoplanets with Deep Learning. V. Improved Light Curve Classification for TESS Full Frame Image Observations

EVAN TEY^{*},¹ DAN MOLDOVAN^{*},² MICHELLE KUNIMOTO,¹ CHELSEA X. HUANG,³ AVI SHPORER,¹ TANSU DAYLAN,^{1, 4, 5} DANIEL MUTHUKRISHNA,¹ ANDREW VANDERBURG,¹ ANNE DATTILO,⁶ GEORGE R. RICKER,⁷ AND S. SEAGER^{7, 8, 9}

¹Department of Physics and Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA, 02139, USA*

²Google *

³University of Southern Queensland, Centre for Astrophysics, West Street, Toowoomba, QLD 4350 Australia

⁴Department of Astrophysical Sciences, Princeton University, 4 Ivy Lane, Princeton, NJ 08544

⁵LSSTC Catalyst Fellow

⁶Department of Astronomy and Astrophysics, University of California, Santa Cruz, CA 95064, USA

⁷Department of Physics and Kavli Institute for Astrophysics and Space Science, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA, 02139, USA

⁸Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA, 02139, USA

⁹Department of Aeronautics and Astronautics, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

ABSTRACT

The TESS mission produces a large amount of time series data, only a small fraction of which contain detectable exoplanetary transit signals. Deep learning techniques such as neural networks have proved effective at differentiating promising astrophysical eclipsing candidates from other phenomena such as stellar variability and systematic instrumental effects in an efficient, unbiased and sustainable manner. This paper presents a high quality dataset containing light curves from the Primary Mission and 1st Extended Mission full frame images and periodic signals detected via Box Least Squares (Kovács et al. 2002; Hartman 2012). The dataset was curated using a thorough manual review process then used to train a neural network called **Astronet-Triage-v2**. On our test set, for transiting/eclipsing events we achieve a 99.6% recall (true positives over all data with positive labels) at a precision of 75.7% (true positives over all predicted positives). Since 90% of our training data is from the Primary Mission, we also test our ability to generalize on held-out 1st Extended Mission data. Here, we find an area under the precision-recall curve of 0.965, a 4% improvement over **Astronet-Triage** (Yu et al. 2019). On the TESS Object of Interest (TOI) Catalog through April 2022, a shortlist of planets and planet candidates, **Astronet-Triage-v2** is able to recover 3577 out of 4140 TOIs, while **Astronet-Triage** only recovers 3349 targets at an equal level of precision. In other words, upgrading to **Astronet-Triage-v2** helps save at least 200 planet candidates from being lost. The new model is currently used for planet candidate triage in the Quick-Look Pipeline (Huang et al. 2020a,b; Kunimoto et al. 2021).

Keywords: Neural networks, Transit photometry, Exoplanet detection methods, Exoplanet Catalogs

1. INTRODUCTION

For three decades, human judgement has played a critical role in the exoplanet revolution that has yielded the discovery of more than 5000 planets outside of the Solar System¹. Exoplanets are typically much cooler, smaller, and fainter than their host stars, so detecting them usu-

ally requires extremely precise observations. At the level of sensitivity required to detect exoplanets, numerous other systematic effects can be present in data that can mimic planetary signals. Separating out these “false positive” signals from true exoplanets has been a major challenge (Jacob 1855; van de Kamp 1963; Bailes et al. 1991) since before the discovery of the first exoplanets in the 1980s and 1990s (Campbell et al. 1988; Latham et al. 1989; Wolszczan & Frail 1992; Mayor & Queloz 1995). Historically, classifying possible planet signals

* These authors contributed equally to the manuscript.

¹ NASA Exoplanet Archive: exoplanetarchive.ipac.caltech.edu

as either false positives or viable planet candidates has most often been carried out by a human inspecting and making a judgement on each signal. Humans are quite well suited for this type of work; we can learn how to distinguish planet candidates and false positives with high accuracy, even after looking at a relatively small number of examples, and often without the benefit of *a priori* knowledge of the “ground truth” of any signal’s true classification.

However, relying on human judgement to separate viable planet candidates from false positives has two main disadvantages. First, humans are slow, both in terms of training time and actual classifications. It often takes months or years of practice for a human to become adept at classifying planets and false positives, and once fully trained, it may take an experienced human several minutes to review all of the information needed to make one classification. At these speeds, even classifying a modest number of possible planet signals ($\sim 10^2 - 10^3$) may take days. Given the rapid increase in the volume of astronomical data available for analysis, it will soon be impractical to rely on human classifications to identify viable planet candidates. Second, humans are inconsistent. Differences in external factors (mood, fatigue, hunger, etc) may cause a human to judge the same signal differently on two different occasions. This makes characterizing and quantifying the biases introduced by human classification challenging and inexact. An alternative system capable of quickly, accurately, and repeatably identifying planet candidates would be highly attractive to planet hunters.

In this paper, we focus on improving a deep neural network classifier used to identify viable planet candidates in data from the *Transiting Exoplanet Survey Satellite* (*TESS*) mission (Ricker *et al.* 2015). *TESS* identifies exoplanets by searching for “transits,” or slight periodic dimmings of the apparent brightness of a star as its planet passes between the star and our vantage point in the Solar System. Transit surveys like *TESS* produce copious numbers ($\gtrsim 10^6$ so far) of false positive signals that must be separated from viable planet candidates to enable discoveries.

Machine learning has become a popular tool for identifying promising planet candidates from transiting exoplanets. Some work has focused on using machine learning to perform the actual planet detection (Pearson *et al.* 2018; Zucker & Giryes 2018; Cui *et al.* 2021), but more often, efforts have focused on using machine learning to classify the large number of possible transit-like signals returned by existing planet detection pipelines. A push early in the Kepler mission (Koch *et al.* 2010; Borucki *et al.* 2010) led to the development of two automated

systems: a decision tree called the Robovetter (Coughlin *et al.* 2016; Thompson *et al.* 2018) and a random forest classifier called the Autovetter (McCauley *et al.* 2015). In that initial work, the Robovetter proved more robust and easily extensible to new regimes and datasets, and therefore was used in the production of fully automated planet candidate catalogs from the Kepler mission.

More recently, Shallue & Vanderburg (2018) introduced a convolutional neural network for vetting planet candidates from the Kepler mission called **Astronet**. Since then, **Astronet** and other similar architectures have been demonstrated on other datasets like K2 (Dattilo *et al.* 2019), TESS (Yu *et al.* 2019; Osborn *et al.* 2020), WASP (Schanche *et al.* 2019), and NGTS (Armstrong *et al.* 2018; Chaushev *et al.* 2019). New tweaks to the methodology including new input information and tweaks to the data representation (Ansdel *et al.* 2018; Jara-Maldonado *et al.* 2020; Valizadegan *et al.* 2021) have yielded improvements in classification performance.

Our work is largely based upon the convolutional neural network originally introduced by Shallue & Vanderburg (2018) and adapted to *TESS* by Yu *et al.* (2019), known as **Astronet-Triage**. Starting in 2019, **Astronet-Triage** had been used in the *TESS* Quick-Look Pipeline (Guerrero *et al.* 2021) to triage planet candidates and remove clear false positives. However, our internal tests revealed that this step resulted in the loss of a fairly large number of viable planet candidates (i.e., “false negatives”). This paper describes our work to improve the performance of **Astronet-Triage** by introducing **Astronet-Triage-v2** to reduce the number of lost planet candidates while throwing out a higher number of false positives.

Our paper is organized as follows: In Section 2, we describe the input transit signals and corresponding light curves which were used for training and testing our classifier, and the labels assigned to each signal. In Section 3, we describe how we processed the data before it is input to our neural network classifier. In Section 4, we describe the architecture of the neural network and the training process. We quantify and present the results of our classifier in Section 5, and we discuss the implications of these results in Section 6. Finally, we conclude in Section 7.

2. DATA

For training and testing our model, we use approximately 25000 human vetted transit signals detected by

the Quick-Look Pipeline (QLP, Huang et al. 2020a,b; Kunimoto et al. 2021) across Sectors 1 – 39.²

2.1. TCEs from TESS FFIs

During its Prime Mission (2018 July 25 – 2020 July 04), TESS collected full-frame images (FFIs) every 30 minutes for 2 years covering 70% of the entire sky (Guererro et al. 2021). The FFI cadence was updated to 10 minutes for the 1st Extended Mission (2020 July 04 – 2022 September 01). QLP produces light curves from these images for all observed targets in the TESS Input Catalog (TIC; Stassun et al. 2018, 2019; Paegert et al. 2021) with TESS-band magnitude (T) brighter than 13.5. Flux time series (raw light curves) from five different sized circular apertures are extracted for each star.

These raw light curves are then filtered to remove low-frequency variability originating from stellar activity or instrument noise. Primarily, this is done by dividing the light curve from each separate orbit by a basis spline (following Vanderburg & Johnson 2014) fit using a break-point spacing between 0.3 days and 1.5 days, selected as described by Shallue & Vanderburg (2018). Finally, these detrended light curves are merged with previous TESS sectors using a shared median value. At this point, an optimal aperture is selected for target star based on its TESS magnitude – fainter stars getting smaller aperture sizes. All subsequent processes use these multi-sector “best”-aperture detrended light curves.

QLP searches these light curves for transit signals using the Box Least Squares (BLS) algorithm (Kovács et al. 2002; Hartman 2012). Because BLS spectra feature a rising trend towards lower frequencies (longer periods), QLP subtracts the low frequency baseline before selecting the highest peak as the detection. For each detected signal, the BLS implementation computes characteristic parameters (orbital period, transit center, transit depth, the full transit duration) by performing a least square trapezoid fit for the transit. These parameters are used later in the input process for **Astronet-Triage-v2**.

Transit signals with signal-to-pink-noise > 9 and BLS peak significance > 5 (for stars with $T < 12$ mag) or > 9 (for stars with $T > 12$ mag) are labelled threshold-crossing events (TCEs). These filters give slightly different perspectives on transit significance: (1) signal-to-pink-noise compares the transit depth to pink noise in the light curve (Pont et al. 2006), while (2) BLS peak significance compares the BLS spectrum’s peak height

to its noise. In combination, these checks help filter out events that are clearly not transit-like.

In addition, we filter out instances where the planet would orbit “inside the star.” For each signal we compute the expected semi-major axis to stellar radius ratio assuming a Keplerian orbit.³ If the ratio < 1 , the signal is labeled as inside the star. Typically, these signals signify stellar variability or blended signals from a smaller nearby star.

2.2. Assembling a set of signals to label

Even with filters described in the previous subsection, manually labeling every TCE would take an enormous amount of time, so we select a subset of TCEs for training / testing. Over time, we gradually accumulated three batches of labeled TCEs from the first two years of TESS Primary Mission (observed with 30 min cadence) and the first year of the TESS 1st Extended Mission (observed with 10 min cadence).

The year 1 (Y1) TESS observations for the southern hemisphere went through significant changes in noise property due to the spacecraft pointing strategy change in Sector 4,⁴ and the subsequent tweaking of the momentum dump frequency. We selected 8992 TCEs detected in Sector 13 (the last sector of Y1) for the labeling. This was not an intentional choice, but after spending hundreds of person-hours labeling these TCEs, we opted to make use of them regardless. Fortunately, despite the fact that our Y1 TCEs came only from Sector 13, the observations that led to these detections still included a diversity of spacecraft pointing control strategies and data artifacts (for example detector warmups following instrument anomaly events⁵). In particular, stars observed in Sector 13 have been observed by TESS in Y1 between one to thirteen sectors and cover a variety of prior sectors.

For the year 2 (Y2) TESS observations in the northern hemisphere, the data has more uniform characteristics including a consistent momentum dump frequency of every 4.4 days starting in Sector 14⁶. We sorted TCEs by their target’s TESS magnitude, and then took the 13372 brightest TCEs detected from Sectors 14–26.

³ When computing the semi-major axis we use two times the detected BLS period in case the detected period is half the true period, which often happens for eclipsing binaries. If the star has an estimate for its mass in the TIC, we use that value; if not, we assume a mass of $1 M_{\odot}$. We also assume a circular orbit.

⁴ https://archive.stsci.edu/missions/tess/doc/tess.drn/tess_sector_04_drn05_v04.pdf

⁵ https://archive.stsci.edu/missions/tess/doc/tess.drn/tess_sector_08_drn10_v02.pdf

⁶ https://archive.stsci.edu/missions/tess/doc/tess.drn/tess_sector_14_drn19_v02.pdf

² QLP data can be found at [doi:10.17909/t9-r086-e880](https://doi.org/10.17909/t9-r086-e880)

In year 3 (Y3), TESS returned to observe the southern hemisphere, with faster cadence and a further improved momentum dump strategy (only once each orbit)⁷. We added an additional 2588 TCEs from Sectors 27-39, which increased the sky coverage and brightness range for our southern hemisphere labels.

We note that TCEs around stars only observed in one of the CCDs in Sector 13 Camera 1, and Camera 1 and 2 for Sector 24 and 25 are not included in our sample due to temporary unavailability of the data at the time of vetting.

Altogether, these TCEs create a broad sample of transit-like events detected in the first three years of TESS observation. The final TCE distribution across the sky is shown in Figure 1, and across *TESS* magnitude (T_{mag}) in Figure 2. Due to the different selection criteria of the TCEs from three different years, they have somewhat different data characteristics. As discussed in Section §5.2, these differences do not significantly impact our results.

2.3. Labels and their definitions

For each TCE we assigned one of the following five labels:

- **E** denotes a *periodic eclipsing signal*. This includes both planetary transits and non-contact eclipsing binaries. In the triage process, we do not take into account information that would distinguish an eclipsing signal from background stars from an eclipsing signal on the target star. Both cases would be labeled as E if they satisfy all the other criteria.
- **S** denotes events containing only a *single transit* or events where an *incorrect period* or *period alias* is assessed to be reported from BLS.
- **B** denotes *contact eclipsing binaries*. They are distinguishable from non-contact binaries through their continuous ingress/egress slope.
- **J** denotes *junk*. This includes other astrophysical phenomena like stellar variability as well as instrumental phenomena like scattered light (due to the Earth or the Moon approaching the field of view and reflecting light into the camera) or artifacts introduced at the times of spacecraft momentum dumps (when the spacecraft’s reaction wheels correct for the spacecraft’s speed).

⁷ https://archive.stsci.edu/missions/tess/doc/tess_drn/tess_sector_27_drn38_v02.pdf

- **N** denotes *not sure*. No conclusive label decision could be made for these TCEs. Often an N label was given when a weak signal bordered on being an E or J.

These labels are not necessarily mutually exclusive. We detail the rules we use in labeling when resolving marginal/ambiguous cases:

- **E vs S**: If there is ambiguity in the period (e.g. both the reported period and the double period are consistent with the data) or the period is only slightly off, we default to an E label. Only if the period is explicitly incorrect (e.g. there are flat light curve segments during expected transits, or there are multiple regular transits outside of expected transit times) do we choose an S label. If there is only one regular transit outside the expected transit time, i.e. it might represent a secondary eclipse, we use an E label, and if the reported period potentially includes the secondary eclipse, we also use an E label.
- **B vs S**: If we have a contact binary with the incorrect period, we default to a B label.

We choose these labels first because they mirror astrophysical phenomena. This means the labeled TCEs provide good targets for follow-up (e.g. Es will be good candidates for exoplanet and binary star detection). Second, we expect similarities in light curve morphology within a label. This should help our model learn labels more accurately.

For the purposes of finding exoplanets, we are particularly interested in high precision and recall metrics for E labels. S and N labels may also be important candidates for further investigation.

2.4. Labeling process

All TCEs were manually assigned labels based on human-visual representations (see Figure 3) similar to the model input representations described in Section 3. On a weekly basis, batches of targets were independently vetted by 3 – 7 of the authors. At the end of the week, targets with conflicting labels where at least one human chose an E or S were discussed in order to reach a consensus on the target’s final label. If a target had only B, J, or N votes, we assigned weights to each label based on the number of votes. Altogether, this process took over 2 years. We expect the multiplicity of vetters to reduce the number of label errors, giving us a very high-quality dataset.

Table A contains examples of signal data along with individually-assigned labels and their consensus dispo-

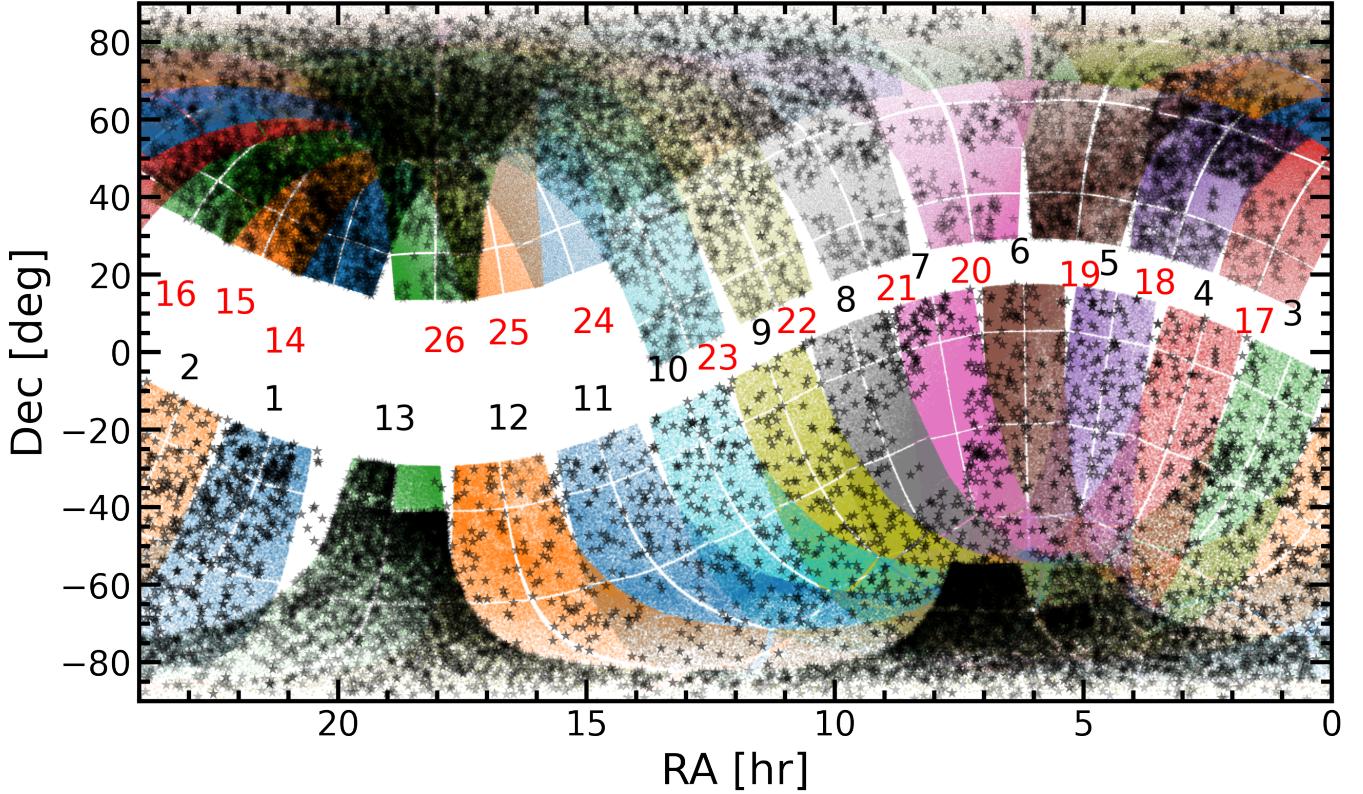


Figure 1. Sky map showing the locations of the 24926 TCEs presented here (black starred data points) compared to the coverage of each TESS Prime Mission sector (colored data points). The black and red labels are the Prime Mission sector numbers in the southern and northern ecliptic hemispheres, respectively. Note that we also include 2588 TCEs from the 1st Extended Mission, for which sector coverage is not shown here. The under- and over-densities of TCEs are due to the selection criteria as described in the text.

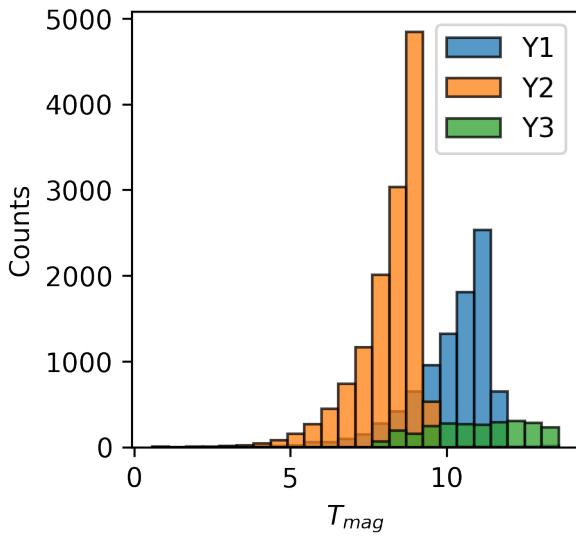
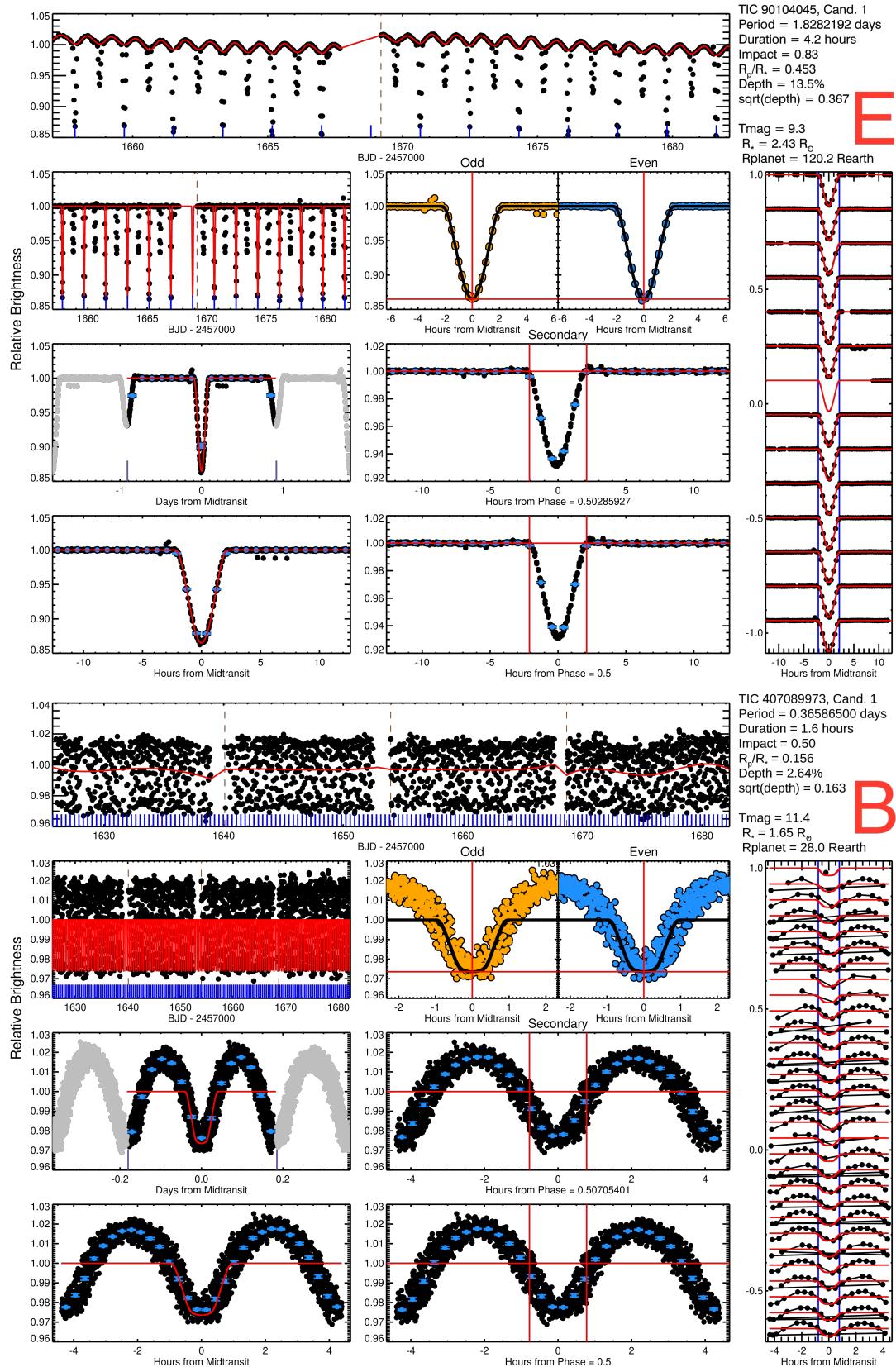


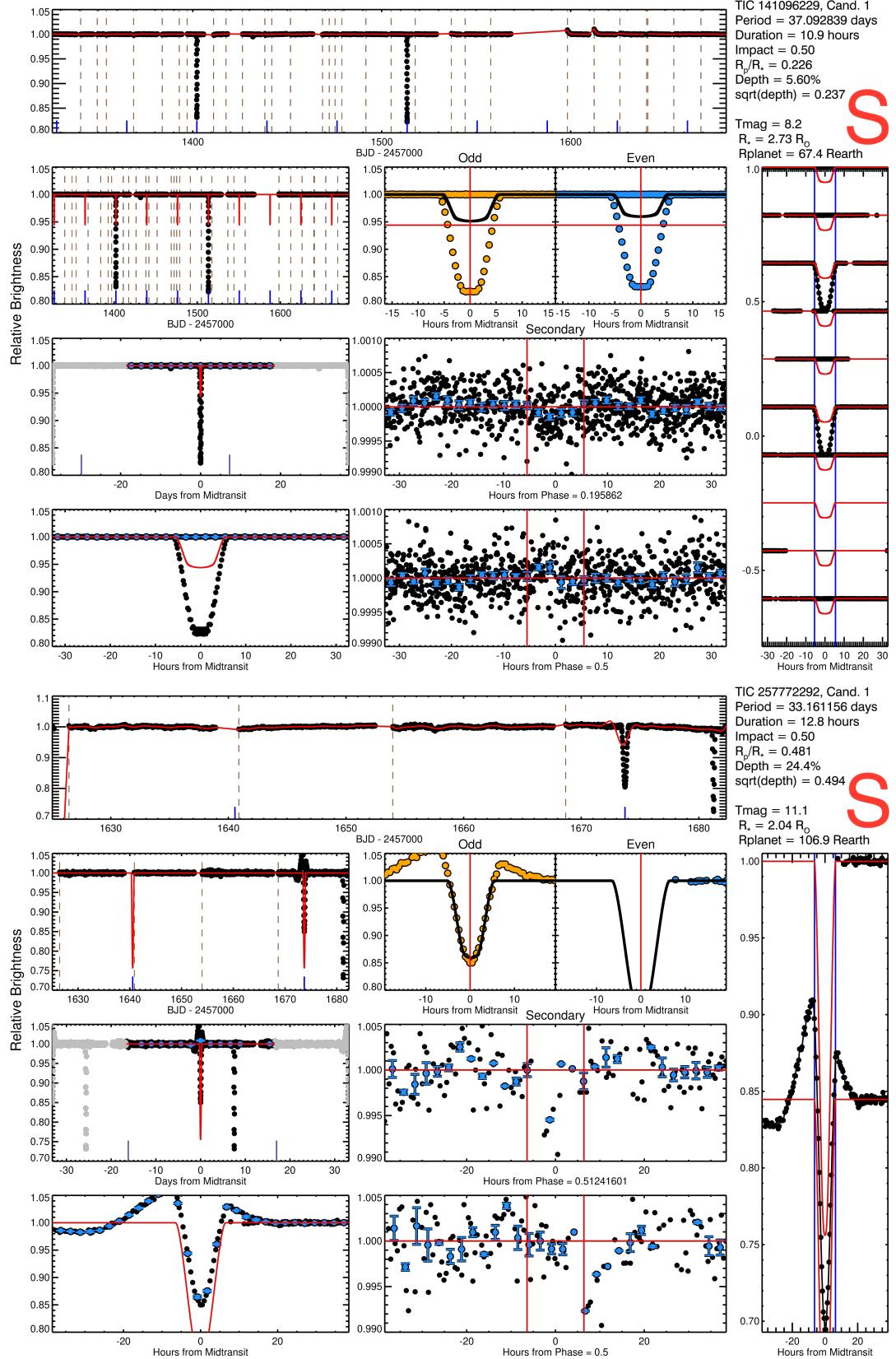
Figure 2. Distribution of T_{mag} across our dataset. Both Y1 and Y2 portions of the dataset focused on the brightest TCEs, while Y3 added TCEs more uniformly across magnitudes. More details on TCE selection can be found in Section 2.1.

sitions. The full table (and accompanying light curve data) can be found online in [Tey et al. \(2022\)](#).

Following common practice in ML, we randomly separate the dataset into a training, validation, and test set. The model is initially fit on the training set, a set of examples used to fit the parameters of the model. Next, the validation set provides a measure of predictive accuracy and model fit. The validation set consists of examples that the model has not seen in the training set and allows for optimization of the architecture and hyperparameters. Lastly, after the model architecture and hyperparameters are finalized, the test set is used as one last objective test of the model accuracy and fit.

1. Training set (**19919 targets**): used for model training. (15414 J + 2102 E + 1681 B + 224 S + 498 N)
2. Validation set (**2491 targets**): used to calculate precision, recall, detection threshold for binary classification, and model debugging. (1945 J + 261 E + 198 B + 17 S + 70 N)





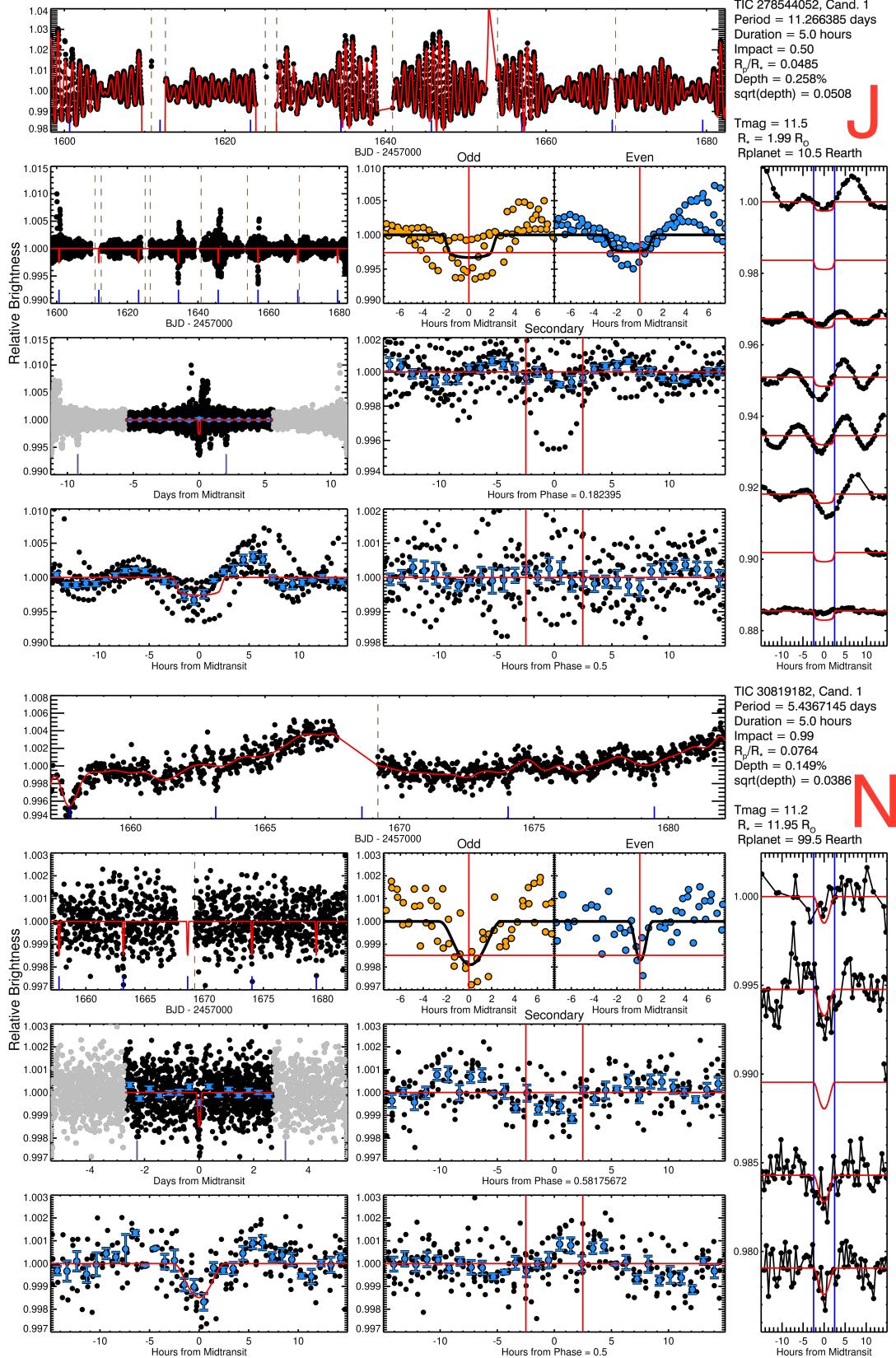


Figure 3. Six example visual representations used for human labeling with labels in red. The different figures within each representation were made to mirror the information described in Section 3. Each image was individually labeled by at least 3 individual vетters. Conflicting labels were discussed and resolved each week.

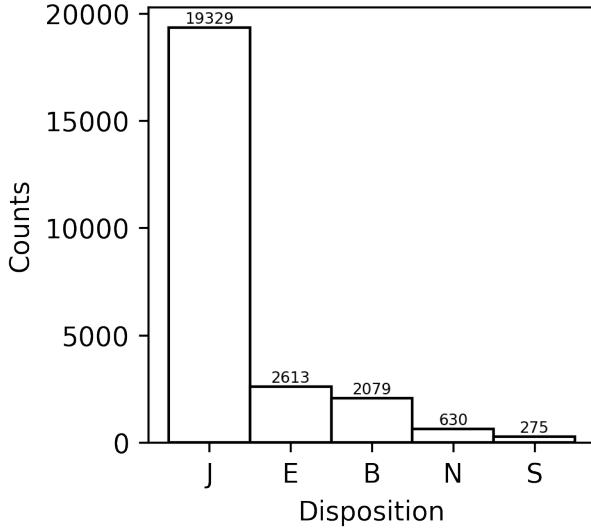


Figure 4. Distribution of labels across our dataset (see Section 2.3 for descriptions of each type). As described in Section 2.4, some TCEs were assigned fractional B and J labels so these counts have been rounded to the nearest integer.

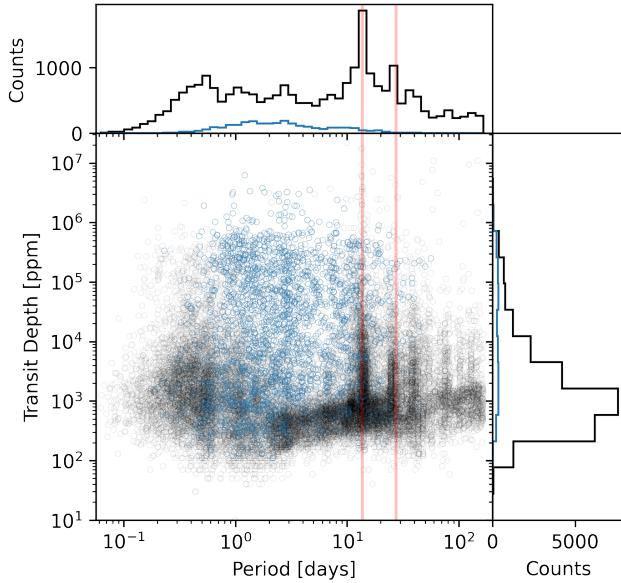


Figure 5. Scatterplot of transit depth vs. orbital period for our dataset. TCEs with E labels are shown in blue. Red lines mark 13.7 and 27.4, the orbital period and twice the orbital period of TESS.

3. Test set (**2516 targets**): hold-out set used for final evaluation; this set was never used for training or debugging, or any other evaluation. (1970 J + 250 E + 200 B + 34 S + 62 N)

2.5. Distribution of the labels

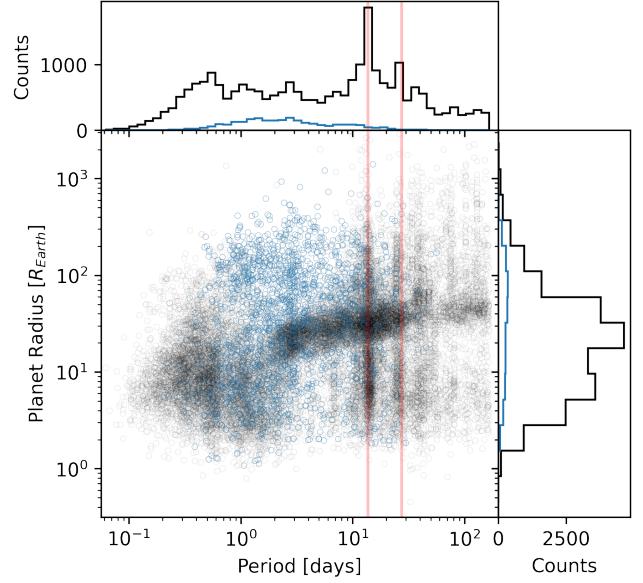


Figure 6. Scatterplot of planet radii vs. orbital period for our dataset. TCEs with E labels are shown in blue. Red lines mark 13.7 and 27.4, the orbital period and twice the orbital period of TESS.

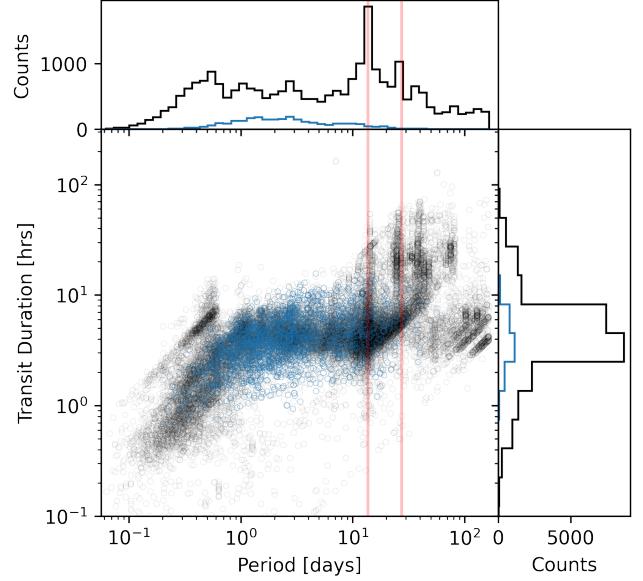


Figure 7. Scatterplot of transit duration vs. orbital period for our dataset. TCEs with E labels are shown in blue. Red lines mark 13.7 and 27.4, the orbital period and twice the orbital period of TESS.

Figure 4 shows the distribution of labels in our training set. Out of the total 24926 labels, the majority are **J** labels (19329). The amount of signals identified as eclipsing objects (**E**, 2613) is comparable to that identified by contact binaries (**B**, 2079).

We examine the distribution of the fundamental transit parameters (i.e., orbital period, transit depth, estimated planet radius, and transit duration) of the labels in Figure 5, 6, and 7. Specifically, we compare the parameter spaces resided by the E labels to the other labels. The comparison reveals the following characteristics: (1) a majority number of the TCEs with period smaller than ~ 0.5 days are not caused by eclipses; (2) a majority of the shallow events with period longer than 10 days are not caused by eclipses; (3) there is clear pile-up of TCEs at the TESS orbital period and its alias, which are not caused by eclipses; (4) a majority of TCEs with extremely short/long transit duration are not caused by eclipses.

3. MODEL INPUT REPRESENTATIONS

For each TCE, we pass the raw flux time series leading to the detection and all the relevant information describing the detected periodic signal and target star to the neural network.

3.1. Time series data

We preprocess the raw flux time series into different input representations before passing them to **Astronet-Triage-v2**. We use the same basis spline techniques used in QLP, however, the transit signals are masked out based on the BLS-detected period, epoch and duration before the optimal spline is computed. This approach will often prevent over-fitting of the transit signals during the detrending process. To account for different time scales of the stellar variability, we adopt multiple detrending settings to provide **Astronet-Triage-v2** a more complete view of the light curve noise characteristics. Unlike in QLP, which only uses one set of splines with spacing between 0.3 and 1.5 days to create the final detrended light curves, we use three different settings (0.3, 5.0, and a value which minimizes the Bayesian Information Criterion, Schwarz 1978) to create three different sets of detrended light curves. The light curves detrended with larger spacing are also less likely to over-fit the transit signals with long transit duration.

For each detrended light curve we generate seven different plots or views (see Figure 8). Each view is binned using a robust binning technique to de-weight outliers. During this binning, we also account for the change in exposure time between the Primary and 1st Extended Mission by weighing points according to their exposure time in a given bin. After this, we normalize the binned data so that the minimum value is -1 and the median value is 0. The complete list of views can be found in

the source code⁸. A detailed description of each view type is below:

- Global View: The global view uses the full light curve folded on the reported period with 201 bins. In addition to the median values, the view also includes the standard deviations for each bin, a mask indicating whether the bin was empty, and a mask indicating whether the bin falls inside the detected transit.
- Local View: The local view uses points within two transit durations of the transit center (for a full timespan of four transit durations), again folded on the reported period. The local view uses 61 bins, and includes standard deviation and mask values like the global view. In addition, we also record the scale factor used in normalization, as a scalar feature.
- Secondary View: The secondary view is similar to the local view, but is centered around the most significant secondary transit, determined by performing a grid search⁹ on the out-of-transit portion of the phase folded view, for a duration equal to the primary transit duration, and selecting the region with the highest signal/noise ratio. This view is accompanied by two scalar features: the normalization scale factor, and the phase of the secondary transit’s center.
- Local Half-Period View: Similar to the local view, but folded at half the detected period. This view only contains the standard deviation value, since the median value can appear very noisy when folding a transit over a non-transit.
- Global Double Period View: Similar to the global view, but folded at twice the period of the global view.
- Sample Global Segments: This view contains the entire period (similar to the global view), but showing up to 7 of the folds that contain the most points (ties are broken at random). Each fold is accompanied by a mask indicating whether the bin contains any points. If the light curve contains

⁸ https://github.com/mdanatg/Astronet-Triage/blob/e4ec517b175b2a3dfb74cf6c6e3f5273dd8749c7/astronet/astro_cnn_model/configurations.py#L2254

⁹ https://github.com/mdanatg/Astronet-Triage/blob/e4ec517b175b2a3dfb74cf6c6e3f5273dd8749c7/light_curve_util/find_secondary.py#L62

fewer transits, the extra views remain empty. Each fold is independently binned with 201 bins.

- **Sample Local Segments:** Similar to the sample global segments, this view contains the transit center of up to 4 of the folds that contain the most points (ties are broken at random), for a total of 8 folds. Each fold is independently binned with 61 bins.

3.2. Scalar data

We also use scalar values that describe characteristics of the transit, host star and the light curve itself. Transit features include period in days (P), transit duration in days (T_{dur}), transit depth (δ), and the number of full periods observed in the flux-time series (n_{folds}), while host star features include TESS magnitude (T_{mag}), mass in M_{\odot} , and radius in R_{\odot} . The host star features are directly extracted from the TESS Input Catalog v8.2 (Paegert et al. 2021).

For TCEs without stellar radii in the catalog, we perform a rough estimate using a Bayesian estimate of the distance (Bailer-Jones et al. 2021), apparent magnitude (either Gaia G, Bp, and Rp, or Gaia G and 2MASS K if Bp and Rp are unavailable), and color/temperature and color/bolometric corrections from MIST models (Choi et al. 2016). In brief, we estimate the temperature and bolometric correction from either the target’s Bp-Rp or G-K colors, use the bolometric correction to estimate the target’s apparent bolometric magnitude, use the estimated distance to the target to convert to an absolute magnitude, convert to bolometric luminosity, and solve for the stellar radius from the temperature and luminosity via the Stefan Boltzmann Law. In our testing, we were able to determine radii within about 10% of the TIC values when present, and provided radius estimates for ~ 2400 from the ~ 2800 TCEs missing stellar radii in our dataset.

Light curve features include the total number of points. Each scalar value is normalized to be zero mean and unit variance across the dataset, except for n_{folds} which is truncated to a maximum value of 100 and a log-scaled to fit between 0 and 1. In addition, we also include as scalar inputs the detected phase of the secondary eclipse, as well as the calculated scaling factor when normalizing the global, local and secondary views.

4. NEURAL NETWORK ARCHITECTURE

Our model uses a convolutional neural network architecture derived from **Astronet**. The high level architecture is shown in Figure 8.

Each time series feature is grouped together with similar features and then passed through a separate convolutional tower. For example, the global view flux is grouped together with the standard deviation of the global view, so that they form a 2-channel, 1-dimensional image. The structure of a convolutional tower is shown in Figure 9. Each tower consists of convolutional layers with Rectified Linear Unit (ReLU) activation, alternating with pooling layers. The pooling layers aggregate neighboring pixels, in effect increasing the field of view of the subsequent convolutional layer.

The output of each convolutional tower is flattened into a vector shape. The flattened outputs from all towers are concatenated together with the auxiliary inputs to form the input into the next section of the network, the fully-connected tower, whose structure is shown in Figure 10. The fully-connected tower is composed of several fully-connected neural network layers, alternating with dropout layers. The dropout layers randomly set inputs to zero, and serve a role of regularization, to mitigate over-fitting. The dropout layers are only active during training. The final layer has five outputs, and uses a sigmoid activation function, so that its output is in the interval [0..1]. Each of the five outputs corresponds to one of the five labels.

The various hyper-parameters of each network can be found in the configuration file included with the source code.¹⁰ The hyper-parameters are tuned using Vizier (Golovin et al. 2017a; Song et al. 2022) by minimizing the loss on the validation set.

4.1. Training

We train the model using the Adam, a popular variant of stochastic gradient descent optimization (Kingma & Ba 2014), for 20,000 steps. The complete set of training parameters can be found in the code¹¹.

For the loss function we use binary cross-entropy loss¹². Notably, this means that the model is not trained to choose between the five labels exclusively. Instead, it produces independent scores for each label, so a model is free to assign high scores for both “E” and “J” labels, for instance. This loss function enables us to assign weighted labels to uncertain examples (e.g. 50 percent

¹⁰ https://github.com/mdanatg/Astronet-Triage/blob/e4ec517b175b2a3dfb74cf6c6e3f5273dd8749c7/astronet/astro_cnn_model/configurations.py

¹¹ https://github.com/mdanatg/Astronet-Triage/blob/e4ec517b175b2a3dfb74cf6c6e3f5273dd8749c7/astronet/astro_cnn_model/configurations.py#L2254

¹² See https://www.tensorflow.org/api_docs/python/tf/keras/losses/BinaryCrossentropy for the implementation and Good (1952) and Shallue & Vanderburg (2018) for more information

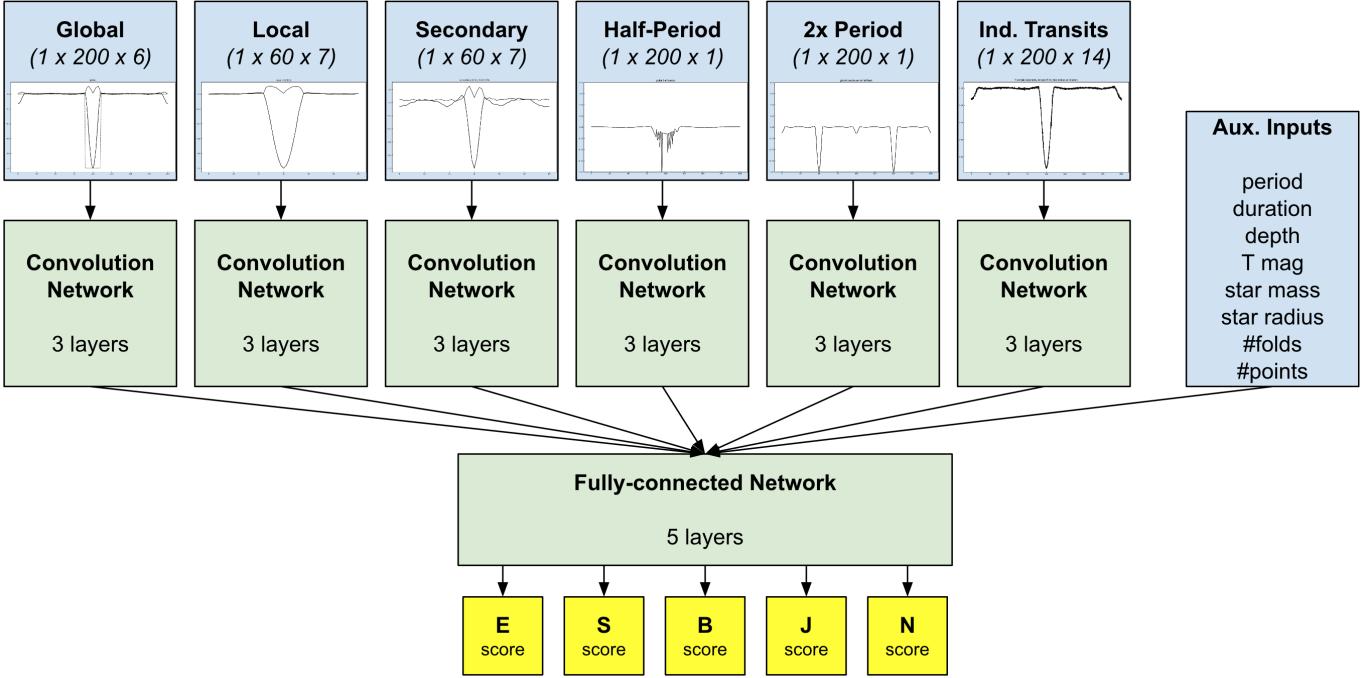


Figure 8. Astronet-Triage-v2 neural network architecture.

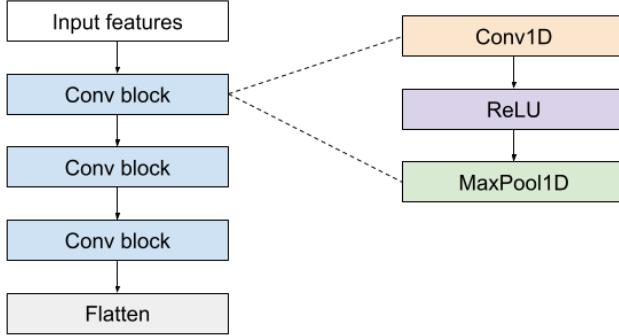


Figure 9. Structure of a CNN tower. Each convolution tower has 1 to 4 blocks. Each block has 1 to 4 layers.

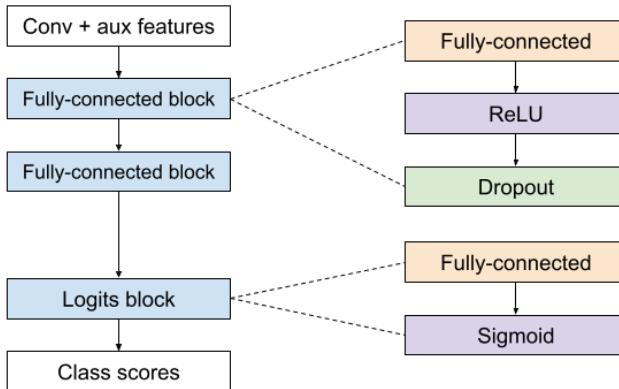


Figure 10. Structure of the fully-connected tower.

“B”, 50 percent “J”). The weight is determined as follows: if a target had a single label (as resulting from the group resolution, or if the vote was unanimous), the weight is 1.0; if the target had multiple votes, the weight is the maximum number of votes for any label divided by the total number of votes. This means targets for which a label didn’t receive a majority of votes are weighted less.

We don’t apply data augmentation, although that is something we intend to do in future work (see Section 6.4.2).

4.2. Prediction and ensembling

As a multi-class classifier, our model outputs a prediction score for each label. Predictions where the “E” label score exceeds a threshold chosen beforehand are considered to predict the label “E”. Otherwise, the model is considered to predict the label with the highest prediction score.

We then construct an ensemble of 10 models trained separately (hence with different initial weight values, and different shuffling of the input data). The compound prediction of the ensemble is constructed as follows:

1. If any of the models predicts “E”, then the ensemble prediction is also “E”.
2. Otherwise, the ensemble prediction is the label predicted by a majority of models, with ties broken at random.

Although the model predicts five different labels, we are primarily interested in the “E” label. The other labels are mainly used at training, to encourage the network to learn natural representations. We found that the extra labels greatly help understand a model’s predictions, as well as validate whether the model does indeed create correct internal representations.

5. RESULTS

Here we report the results of our ML activity predictions. First we discuss the metrics we used to evaluate the performance and then we summarize how the different models performed on each dataset.

The two primary metrics we use to evaluate our performance are precision and recall. The precision, or reliability, of a model on a labelled dataset is the number of true positives divided by the number of true positives and false positives. Recall, or completeness, is the number of true positives divided by the number of true positives and false negatives. As we are interested in “E” labels as potential planet candidates, they generally are used as the “positive” class. In this context, a high precision means our model outputs fewer false positives, meanwhile a high recall means successful recovery of more planet candidates (fewer potential planets lost by **Astronet-Triage-v2**). Since labels are determined by comparing output prediction scores against a chosen threshold, each specific threshold yields its own precision and recall. When plotted over many different thresholds, we can form a precision-recall curve (see Figure 11). By taking the area under the precision-recall curve (AUC-PR), also known as the average precision, we can characterize our model’s overall performance and compare against other models with the highest achievable value being a 1.

5.1. Performance on validation and test sets

On the validation dataset we obtained an AUC-PR value of 0.977. The model achieves 100% recall at 41% precision, at a prediction threshold of 0.0105. If we increase the threshold to 0.215, we obtain 96.9% recall at 79.8% precision.

On the test set, we obtained an AUC-PR value of 0.965. The model achieves 100% recall at 15% precision, at a prediction threshold of 0.0005. This suggests the test set contains more difficult examples (possibly incorrect ones). With the thresholds suggested by the validation set, we obtain 99.6% recall at 39.7% precision for the 0.0105 threshold, and respectively 97.2% recall at 75.7% precision for the 0.215 threshold.

5.2. Generalizing to TESS 1st Extended Mission data

We explore the adaptability of our network, and the generalization of training on non-uniform datasets in this section. In practice, models like **Astronet-Triage-v2** are trained on previously observed sectors with a goal of classifying new observations taken by TESS in the future. Since noise characteristics and TESS observation strategy can change sector-to-sector, it is important that our models generalize well to new data.

Nearly 90% of our total training dataset comes from the TESS Primary Mission, so we use QLP data from TESS 1st Extended Mission (Sector 33, observed during Year 3 from UT 2020 December 17 – UT 2021 January 13) to test how our model generalizes to unseen or out-of-distribution data.

Following the QLP convention, we ran a BLS search and **Astronet-Triage-v2** on the full multi-sector light curves (including both Primary Mission and 1st Extended Mission data) for each star. Of the discovered TCEs, we selected a random sample of 759 targets with $T_{\text{mag}} < 11$ from camera 1 and 590 targets with $11 < T_{\text{mag}} < 13.5$ from camera 2. Due to the TESS pointing strategy, we focus on these cameras because their light curves have roughly equal amounts of Primary vs. 1st Extended Mission observations. The magnitude ranges also allow us to compare performance on stars in different brightness bins.

One of our vetters (CH) independently labeled all 1349 TCEs before evaluation, among which, 255 TCEs were assigned an E label.

To better understand our ability to generalize, we apply the following models to the Sector 33 dataset: **Astronet-Triage**, the fully trained **Astronet-Triage-v2**, and three independent instances of the **Astronet-Triage-v2** architecture trained on different subsets of our original TCE dataset (Section 2).

These three separate training sets were formed by splitting our original training set on observation year, meaning roughly 40% went into training the Y1 model, 50% into the Y2 model, and 10% into the Y3 model. The differences between these datasets are described in Section 2.1, but briefly: Both the Y1 and Y2 datasets feature brighter stars, but the Y1 dataset were only taken from Sector 13, so they cover a small region of the Southern ecliptic hemisphere. The Y2 dataset, on the other hand, were selected more uniformly and cover most of the Northern ecliptic hemisphere. Neither has much overlap in sky coverage with the evaluation set (the 1349 Sector 33 TCEs) – Y1 having little overlap and Y2 having none. Both datasets also have much shorter observation baselines than the evaluation set, and finally, due to the change in TESS momentum dump strategy, the Y1 dataset also differs from the evaluation

Table 1. Performance on previously unseen S33 data

Model	Cam	Threshold	Precision	Recall
Astronet-Triage-v2	1	0.0105	0.64	0.98
Astronet-Triage-v2	2	0.0105	0.53	1.00
Astronet-Triage-v2	1	0.215	0.89	0.91
Astronet-Triage-v2	2	0.215	0.84	0.99
Astronet-Triage	1	0.08	0.89	0.85
Astronet-Triage	2	0.08	0.82	0.90

set in noise characteristics. The Y3 dataset bears the most similarity to the evaluation set in terms of data characteristic. It is, however, much smaller than the other datasets. Altogether, these different datasets and models provide useful views at our ability to generalize to data that can be fairly different from the training data.

Since **Astronet-Triage** only distinguishes between transit-like and non-transit-like, it's trained to give high scores TCEs we consider E- or S- labeled. As **Astronet-Triage-v2** provides independent E and S scores, we choose remove all S-labeled data from precision and recall calculations for a simple direct performance comparison with **Astronet-Triage**. This leaves us with 1315 TCEs.

Precision and recall numbers split across **Astronet-Triage** and **Astronet-Triage-v2** for each camera can be seen in Table 1. In both cameras we see that for similar (or better) levels of precision, **Astronet-Triage-v2** provides better recall than **Astronet-Triage**, with a slightly more pronounced effect in camera 2 (fainter targets). In other words, for the same amount of human vetting time, **Astronet-Triage-v2** would recover more potential planets than **Astronet-Triage**.

The full precision-recall curves across all TCEs (ignoring S-labeled TCEs) are shown in Figure 11. Across the board we see that **Astronet-Triage-v2** (trained on the full training set) improves on **Astronet-Triage** with AUC-PR scores of 0.961 and 0.927. We also see that the models trained only on Y1, Y2, and Y3 data perform similarly to **Astronet-Triage** with AUC-PR scores of 0.954, 0.960, and 0.917 respectively. Even though the Y1 and Y2 versions of the models don't use any 1st Extended Mission training data, we see they're still able to perform highly in S33 (which occurred during Y3). This supports **Astronet-Triage-v2**'s ability to generalize to future sectors.

5.3. Performance on the TOI catalog

The TESS Objects of Interest (TOI) catalog (Guerrero *et al.* 2021), which lists the planetary candidates detected by *TESS*, is a useful benchmark for high-

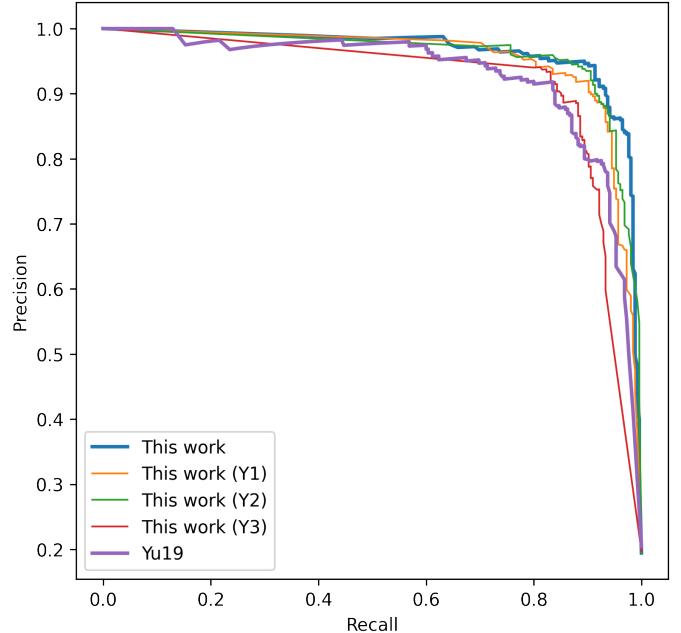


Figure 11. Precision vs. recall for 1315 TCEs selected from Sector 33 of the 1st Extended Mission. Since **Astronet-Triage** (Yu *et al.* 2019) only distinguishes between transit-like and non-transit-like, it gives high scores to TCEs we either consider to have E or S labels. For a more direct comparison to **Astronet-Triage-v2**, we choose to ignore all S-labeled TCEs when calculating precision and recall. We see that across all levels of recall, **Astronet-Triage-v2** provides higher precision even when trained only on Primary Mission data taken during Y1 or Y2. Although the Y3 dataset bears the most resemblance to the S33 evaluation set here, the size of the Y3 dataset is only ~ 2500 , so the Y3-trained model doesn't quite reach the performance of the other models.

confidence E or S labels. A good model should label all TOI entries as E or S, since humans have inspected each entry and considered them to be high-probability planetary candidates (allowing for single-transit events).

On 2022 April 21 we downloaded the TOI catalog with light curve data through Sector 47. We also use information from TESS Follow-up Observing Program (TFOP) Sub Groups 1 and 2 (SG1 & SG2), which use ground-based photometry and reconnaissance spectroscopy to follow-up on TOIs and help filter out false positives. After keeping only planet candidates (PCs; meaning TOIs that were not ruled out as false positives with follow-up observations) and validated / confirmed / known planets (Ps), we have a dataset of 4140 targets.

After evaluating all TOI signals with **Astronet-Triage-v2**, Figure 12 shows the distribution of E scores. Figure 13 shows the recall rate at different cutoff threshold levels. We see that 93% of the TOIs have E scores > 0.0105 and as we increase the cutoff to 0.215,

Astronet-Triage-v2 passes 86% of the TOIs. We also see improved **Astronet-Triage-v2** performance on known, confirmed, or validated planets (Ps) compared to the planet candidates (PCs) across the board.

For comparison, we also ran **Astronet-Triage** on all TOI signals. Using a threshold of 0.09, as was originally used in QLP, **Astronet-Triage** recovers 3349 TOIs. Using the dataset from Section 5.2, we find a precision-matching threshold of 0.2 for **Astronet-Triage-v2**. By finding the threshold of equal precision, we can compare TOI recovery at a constant rate of human vetter work. At this threshold, 3577 TOIs are recovered. In other words, at least 200 TOIs are saved by using **Astronet-Triage-v2** in place of **Astronet-Triage** without introducing more false positives to human veters.

Some important caveats to note:

- The TOI catalog does include single-transit events. **Astronet-Triage-v2** is trained to give these S rather than E labels. Rather than keeping separate cutoffs for S and E scores, for simplicity we choose to focus on E scores in reported recalls. This gives it a slight disadvantage in terms of recovery numbers, though we leave them in the dataset for fairer comparison to **Astronet-Triage** which gives a score for transit-like (periodic or single-transit) versus not transit-like.
- TOIs can also come from the SPOC pipeline, which processes 2-minute cadence light curves. For both **Astronet-Triage-v2** and **Astronet-Triage**, QLP light curves are binned down to 30 or 10 minutes, so some signals may not be detectable (e.g. due to low signal-to-noise in the binned light curve) and should be assigned J labels. This contributes partially to the lower recall numbers seen at the cutoffs from Section 5.1.
- Only 130 TOI host stars appear in our dataset of $\sim 25,000$, 100 of which were in the training set. We also conducted this analysis with those TOIs removed and saw similar results.

6. DISCUSSION

6.1. Use in producing the TOI catalog

A large piece of motivation for this work has been improving on **Astronet-Triage** so fewer planet candidates are lost when searching for TOIs via QLP. After signal detection via BLS, Astronet is one of the finals triage steps before candidates are passed along to human TOI vетters and potentially promoted

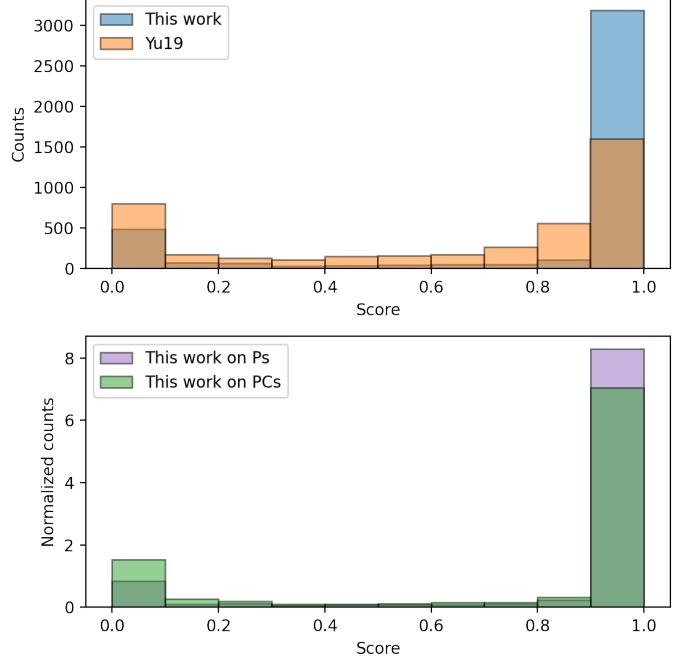


Figure 12. Top: Distribution of E score between this work and **Astronet-Triage** (Yu et al. 2019) on the whole TOI dataset. Bottom: Distribution of E scores from this work when the dataset is separated into Planets (P, validated, confirmed, and known planets) and Planet Candidates (PC, TOIs that are not validated, confirmed, or known planets, and were also not identified as false positives with follow-up observations).

to TOIs (Guerrero et al. 2021). Based on the results in Section 5 we expect **Astronet-Triage-v2** to save many planet candidates that would otherwise be lost without adding false positives and increasing the hours needed for human TOI vetting. Starting in Sector 34, early versions of **Astronet-Triage-v2** officially replaced **Astronet-Triage** within QLP. While **Astronet-Triage-v2** takes step towards a more automated process, it is still not developed enough for population statistics (for a deeper discussion see Section 6.4.1).

6.2. What is limiting our precision?

In our tests, we found a common source of false negatives stemming from patterns with borderline label assessments. The most common being eclipsing binaries which are non-contact but still close enough to resemble the pattern of a contact binary, due to, for example, tidal distortion, hence it is unclear whether the label should be “E” or “B” (Figure 14). Other instances of ambiguous patterns are represented by very noisy transits, or transits on a background of high stellar variabil-

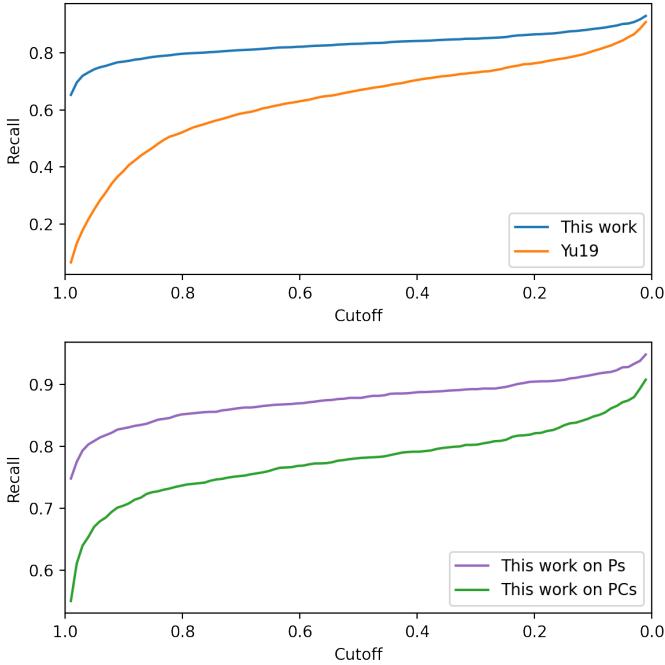


Figure 13. Top: Recall as a function of cutoff threshold between this work and **Astronet-Triage** (Yu et al. 2019). For **Astronet-Triage-v2** we choose to focus on just E scores even though some TOIs are true S labels. Bottom: **Astronet-Triage-v2** recall as a function of cutoff threshold when the dataset is separated into Planets (P, validated, confirmed, and known planets), and Planet Candidates (PC, TOIs that are not validated, confirmed, or known planets, and were also not identified as false positives with follow-up observations).

ity, where the distinction between “E” and “J” is more subtle (Figure 15).

One particular element of sensitivity for the neural network is on the correctness of the period and duration values estimated by BLS. Errors in these values can lead to de-trending distortions which can make phase-folded views deviate from a transit-like light curve shape. Examples containing multi-year observations can be particularly sensitive, as even slight variations in the detected period can lead to a blurring of the transit in the phase folded view (Figure 16 and 17).

We also note that the phase folding and binning processes are inherently lossy (similar to how compressing an image is a lossy process). While we have not ascertained the impact of such loss of information, it is to be expected that it causes some loss of precision.

6.3. Comparison to other works

Our work is largely based on the original TESS **Astronet-Triage** classifier described by Yu et al. (2019), which was used for QLP planet candidate triage

from Sectors 6 to 33. The following summarizes the major differences in development and implementation between classifiers:

1. **Astronet-Triage** was trained and tested on QLP light curves from only TESS Sectors 1 – 5, while **Astronet-Triage-v2** was trained and tested on Sectors 1 – 39.
2. **Astronet-Triage** was developed using 16,516 labeled TCEs (493 planet candidates, 2155 eclipsing binaries, and 13,868 noise/systematic signals), which is roughly two-thirds the size of our labeled set (24,926 TCEs).
3. **Astronet-Triage** used labels that were assigned by only a single vetter who visually inspected all TCEs, while 3 – 5 vetters independently inspected each of the TCEs for **Astronet-Triage-v2**, and group discussions resolved labeling disagreements. As a result, our labels should be more reliable.
4. **Astronet-Triage** only labels signals as either “planet” (for all eclipsing signals, including planets and eclipsing binaries) and “non-planet” (for other false positives, including pulsating variables, noise and systematics). The five-label model used by **Astronet-Triage-v2** (E, S, B, J, N) is more flexible and informative.
5. **Astronet-Triage** takes the light curves already detrended by QLP, and bins the data into two views: a “global” view, showing the full light curve phase diagram, and a “local” view, showing a close-up of the transit in the phase diagram. As described in Section 3, **Astronet-Triage-v2** creates three sets of detrended light curves from the raw QLP light curve, and generates seven views for each one. In total, **Astronet-Triage-v2** uses 21 unique views to inform its classification compared to the two used by **Astronet-Triage**.

These key differences result in improvements to our ability to classify TESS signals in FFI data, as shown Sections 5.2 and 5.3.

To our knowledge Yu et al. (2019) is the only truly comparable work to ours, in that their source dataset was the TESS Full Frame Images and not the pre-selected targets processed by the SPOC pipeline, and, their goal was to perform triage by identifying all eclipsing signals, rather than separating planet candidates from eclipsing binaries and other false positives. Some other groups have trained and tested neural networks on TESS data from two-minute postage stamps processed

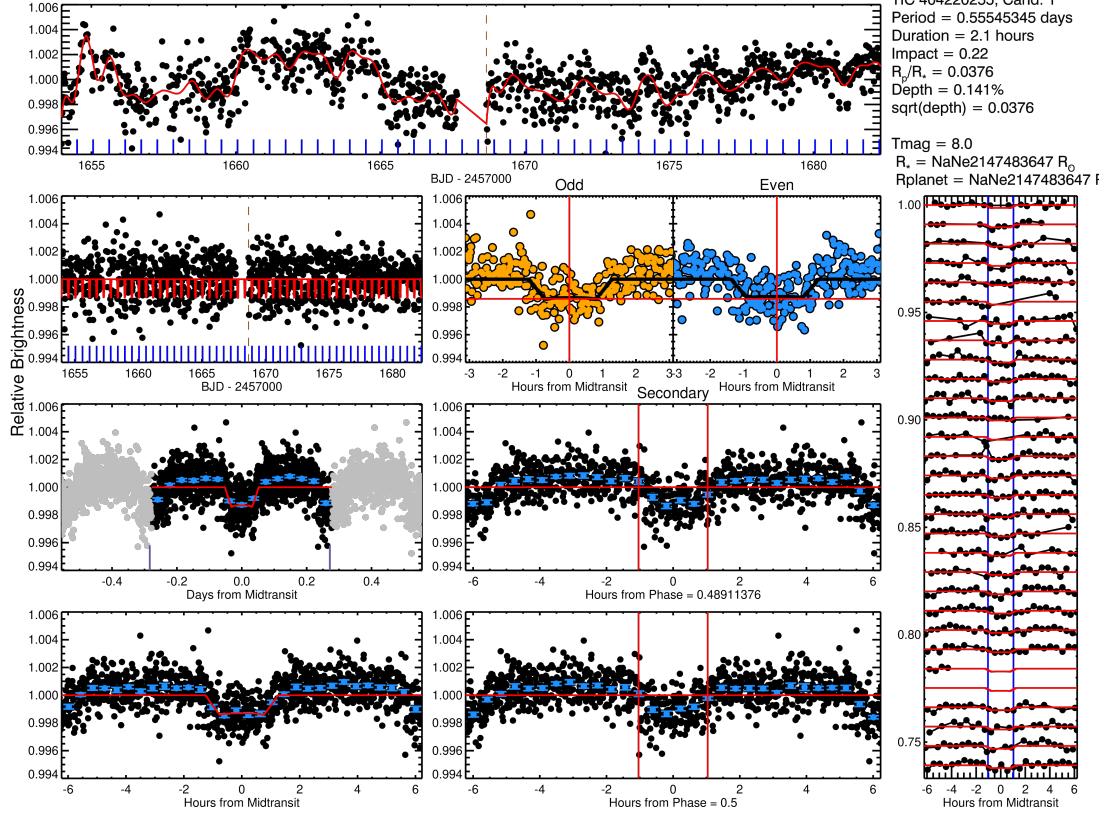


Figure 14. Example of borderline pattern. The true label for this example is “E”, but the folded light curve appears very similar to a “B”.

by the SPOC pipeline (Osborn et al. 2020; Rao et al. 2021; Valizadegan et al. 2021; Fiscale et al. 2021; Ofman et al. 2022), and were successful in identifying planet candidates. However, in general, these groups find that the neural network performance is worse on TESS data than a similar network on Kepler data, likely due to TESS’s higher *a priori* TCE false positive fraction (due to the larger TESS pixels resulting in more blending) and shorter observational baseline. The false positive rate for FFI targets is likely even higher because a) the targets observed by QLP tend to be fainter than targets observed in postage stamps and blending is more pronounced, and b) the targets observed in the FFIs are more often large, luminous stars like red giants, which are difficult to find planets around, and are photometrically noisy. Therefore, TCEs detected by the QLP likely have an even higher *a priori* false positive probability than TCEs detected by TESS in postage stamp data.

6.4. Future work

6.4.1. Applications to exoplanet population statistics

Planet catalogs can be used to characterize exoplanet population statistics through the estimation of occurrence rates. One of the key components of occurrence

rate methodologies is a characterization of catalog completeness, reflecting how many planets from the underlying population were missed. A second key component is an understanding of catalog reliability (Bryson et al. 2020), reflecting how much of the catalog is polluted with false positives. For these reasons, occurrence rate studies require the ability to produce planet catalogs in a fully automated, uniform, and reproducible way, rather than relying on biased manual identification of planet candidates.

NASA’s Kepler mission has dominated the past decade of demographics work in large part thanks to the fully automated Kepler Robovetter pipeline, which enabled careful characterization of both completeness and reliability across wide areas of exoplanet parameter space (Thompson et al. 2018; Christiansen et al. 2020). However, there is not yet a fully automated TESS planet vetting pipeline. Most previous work has also focused on 2-minute cadence observations rather than FFIs, which will be less suitable for demographics due to selection biases in 2-minute cadence target lists. **Astronet-Triage-v2** is an important step toward uniformly vetted FFI planet catalogs, and it naturally allows for a flexibility in balances between completeness and reliability through the adjustment of prediction

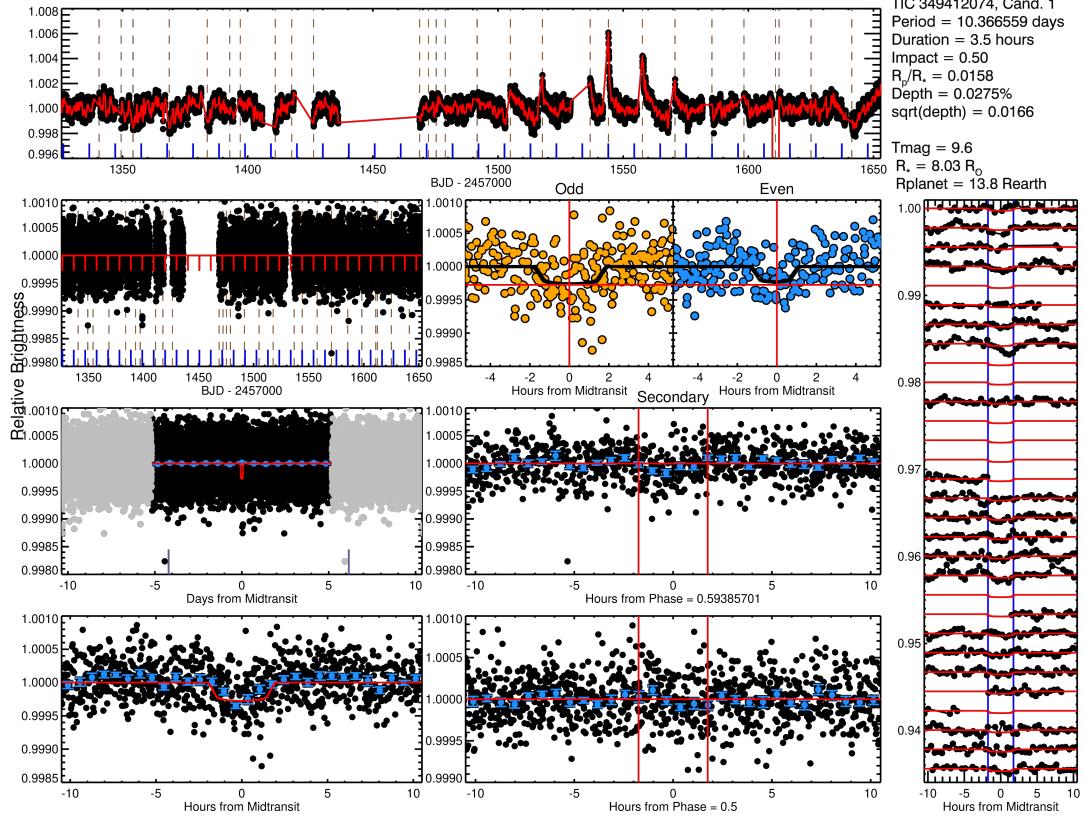


Figure 15. Example of borderline pattern. The very low signal to noise ratio of the transit signal is easily mistaken for a “J”.

thresholds for passing candidates. While the classifier is not yet able to distinguish eclipsing binary false positives from planets (labeling all such signals as “E”’s), it can be used as a first round of automated and characterizable triage. Future improvements to **Astronet-Triage-v2** (Section 6.4.2) are expected to improve the precision and recall, and therefore the completeness and reliability, of any resulting planet catalog. We have plans to extend **Astronet-Triage-v2** to be capable of all steps of the vetting process in the future.

6.4.2. Further improvements to the neural network

In future work, we suggest a number of additions to further improve the performance of our classifier.

Over the past few decades, the performance of deep learning classifiers has seen unprecedented success. A large part of this success has been attributed to the increasing size of training datasets. In this work, the number of training examples is relatively low, particularly for the S-labelled class, with a large class-imbalance (see Figure 4).

A common technique for increasing training datasets, without obtaining new labelled data, is data augmentation. This typically involves applying slight transformations to the training data to produce new data that mimics real observation. Using a combination of a few

data augmentation techniques can magnify a training set by several fold and helps reduce over-fitting. In future work, we suggest applying data augmentation methods such as randomly reversing or clipping light curves in time and applying random Gaussian noise to the light curves or scalar features. We note that these methods were applied in Ansdell *et al.* (2018), where they showed that the main benefit to data augmentation on exoplanet classification was alleviating model over-fitting, with only a small improvement to model performance. More complex augmentation methods such as fitting a model (e.g. Gaussian Process, see Boone 2019) to the minority class light curves and generating more synthetic data may also help to improve the limited data for some classes.

Since **Astronet-Triage-v2** is used in production for QLP’s monthly planet search, another way to increase our training dataset is to use the existing human vetting work that goes into producing the TOI catalog (Guererro *et al.* 2021). As this human vetting is the final step in the TOI release process, there is a high level of quality control in the labels and the signals being vetted are often the most difficult to classify, making them important examples for the model to learn.

7. CONCLUSION

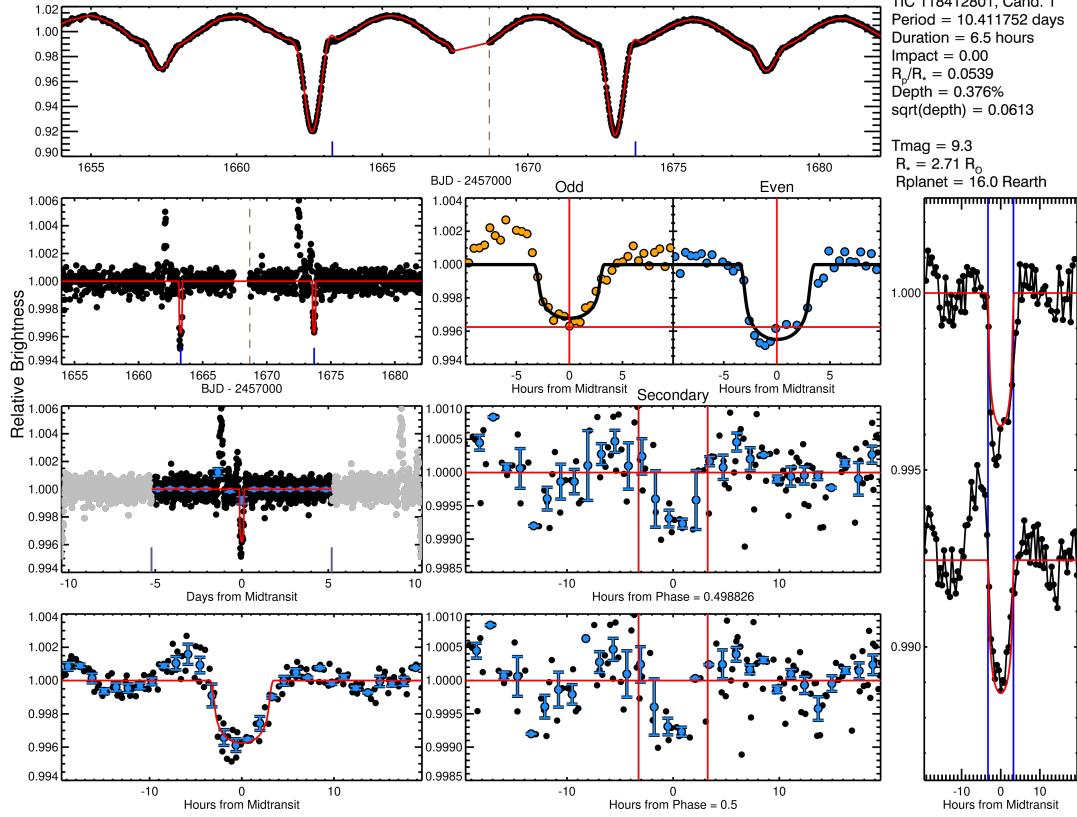


Figure 16. Example of incorrect BLS estimation. Although the phase and period are close, the transit duration is too small, causing the transit to be clipped by the detrending process.

We have presented **Astronet-Triage-v2**, a convolutional neural network designed to distinguish astrophysical eclipsing candidates from other phenomena such as stellar variability and instrumental systematics in TESS FFI light curves. The network assigns input signals one of five labels, namely “E” for eclipsing signals, “S” for single transits or incorrect periods, “B” for contact binaries, “J” for signals due to noise or systematics, and “N” for inconclusive cases. We trained **Astronet-Triage-v2** using ~ 25000 signals, which were detected by QLP from TESS Sectors 1 – 39 and human-labeled through manual review and group discussion. We make this training set available to the community.

Astronet-Triage-v2 is the next in a line of **Astronet** architectures, which were first used for Kepler (Shallue et al. 2019) and later extended to K2 (**Astronet-K2**; Dattilo et al. 2019) and TESS (**Astronet-Triage**; Yu et al. 2019). This iteration features significant improvements over **Astronet-Triage**, including a larger and more robust training set, an expanded list of possible classifications, and more than ten times the number of unique views used to analyze each signal. As a result, we found **Astronet-Triage-v2** is more successful at correctly labeling known TOIs across all

most all cutoff values, with 86% recall at a cutoff of 0.215 compared to 82% recall by **Astronet-Triage**. When tested on a set of new signals from Sector 33, **Astronet-Triage-v2** provides better recall of E and S labels than **Astronet-Triage** for similar (or better) levels of precision, especially for fainter targets. Starting in Sector 34, **Astronet-Triage-v2** officially replaced **Astronet-Triage** within QLP.

As both the TESS observing baseline and number of observed stars continue to increase, automated TESS planet vetting tools will become more important. This is especially true of tools tuned for planet searches using FFIs, of which **Astronet-Triage-v2** is one of the few currently available. While **Astronet-Triage-v2** is not yet capable of distinguishing between eclipsing binaries and transiting planets, it serves as an effective first round of automated and characterizable triage. We plan to continue to improve and extend the network into a fully automated vetting tool in the future.

ACKNOWLEDGEMENTS

This paper includes data collected by the TESS mission. Funding for the TESS mission is provided by the NASA’s Science Mission Directorate.

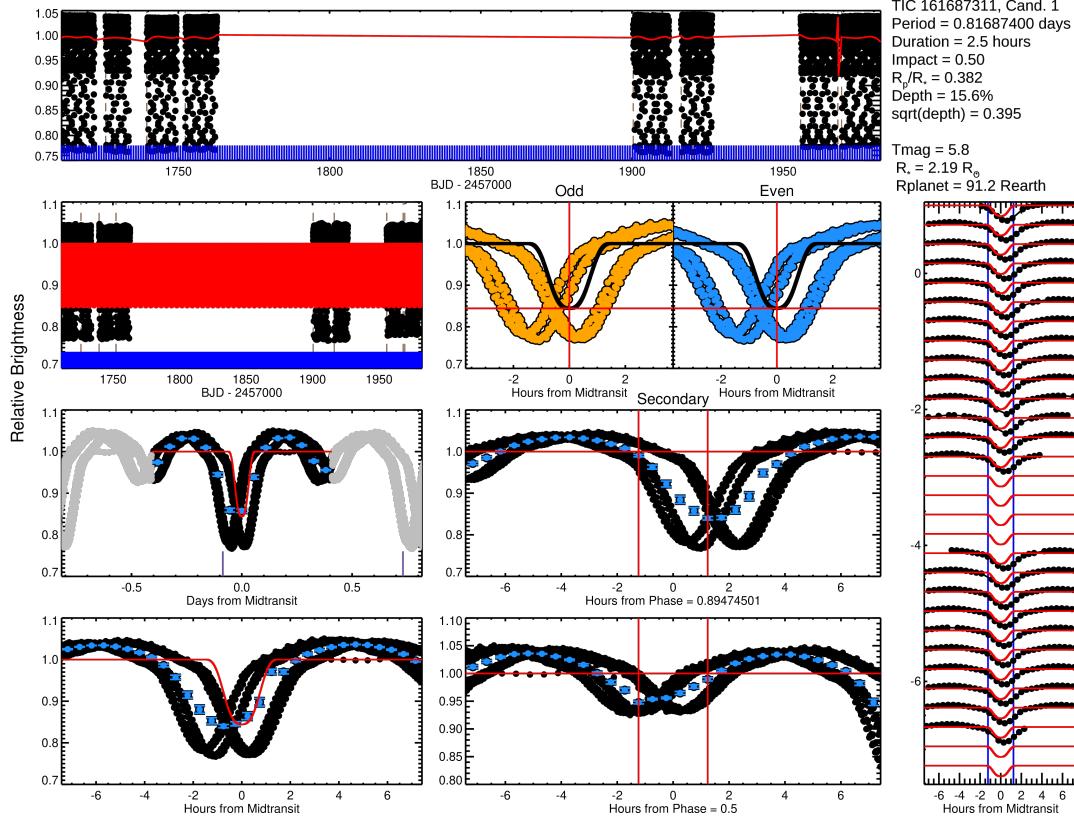


Figure 17. Example of incorrect BLS estimation. The detected period is close, but when the light curve contains a large number of folds, the error compounds and leads to a blurring of the transit view. This is due to QLP searching the light curve with an undersampled BLS frequency grid (necessary due to the computational time needed to run BLS on a large number of targets each sector), as discussed in Kumimoto *et al.* (2022, in prep.).

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

This work was supported by an LSSTC Catalyst Fellowship awarded by LSST Corporation to T.D. with

funding from the John Templeton Foundation grant ID #62192.

The *Astronet-Triage-v2* model was trained and tuned on Google Compute Engine.

Facility: TESS, Gaia

Software: numpy (Oliphant 2006), matplotlib (Hunter 2007), pandas (pandas development team 2020; Wes McKinney 2010), statsmodels (Seabold & Perktold 2010), pydl, astropy (Astropy Collaboration *et al.* 2013; Price-Whelan *et al.* 2018), TensorFlow (Abadi *et al.* 2016), Vizier (Golovin *et al.* 2017b), Jupyter (Kluyver *et al.* 2016)

REFERENCES

- Abadi, M., Agarwal, A., Barham, P., et al. 2016, TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, doi:10.48550/ARXIV.1603.04467
- Ansdell, M., Ioannou, Y., Osborn, H. P., et al. 2018, ApJL, 869, L7
- Armstrong, D. J., Günther, M. N., McCormac, J., et al. 2018, MNRAS, 478, 4225
- Astropy Collaboration, Robitaille, T. P., Tollerud, E. J., et al. 2013, A&A, 558, A33
- Bailer-Jones, C. A. L., Rybizki, J., Fouesneau, M., Demleitner, M., & Andrae, R. 2021, AJ, 161, 147

- Bailes, M., Lyne, A. G., & Shemar, S. L. 1991, *Nature*, 352, 311
- Boone, K. 2019, *AJ*, 158, 257
- Borucki, W. J., Koch, D., Basri, G., et al. 2010, *Science*, 327, 977
- Bryson, S., Coughlin, J. L., Kunimoto, M., & Mullally, S. E. 2020, *AJ*, 160, 200
- Campbell, B., Walker, G. A. H., & Yang, S. 1988, *ApJ*, 331, 902
- Chaushev, A., Raynard, L., Goad, M. R., et al. 2019, *MNRAS*, 488, 5232
- Choi, J., Dotter, A., Conroy, C., et al. 2016, *ApJ*, 823, 102
- Christiansen, J. L., Clarke, B. D., Burke, C. J., et al. 2020, *AJ*, 160, 159
- Coughlin, J. L., Mullally, F., Thompson, S. E., et al. 2016, *ApJS*, 224, 12
- Cui, K., Liu, J., Feng, F., & Liu, J. 2021, arXiv e-prints, arXiv:2108.00670
- Dattilo, A., Vanderburg, A., Shallue, C. J., et al. 2019, *AJ*, 157, 169
- Fiscale, S., Ciaramella, A., Inno, L., et al. 2021, *Research Notes of the American Astronomical Society*, 5, 91
- Golovin, D., Solnik, B., Moitra, S., et al. 2017a, in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Halifax, NS, Canada, August 13 - 17, 2017 (ACM), 1487–1495
- Golovin, D., Solnik, B., Moitra, S., et al. 2017b, in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17 (New York, NY, USA: Association for Computing Machinery), 1487–1495
- Good, I. J. 1952, *Journal of the Royal Statistical Society. Series B (Methodological)*, 14, 107
- Guerrero, N. M., Seager, S., Huang, C. X., et al. 2021, *ApJS*, 254, 39
- Hartman, J. 2012, *VARTOOLS: Light Curve Analysis Program*, Astrophysics Source Code Library, record ascl:1208.016, ascl:1208.016
- Huang, C. X., Vanderburg, A., Pál, A., et al. 2020a, *Research Notes of the American Astronomical Society*, 4, 204
- . 2020b, *Research Notes of the American Astronomical Society*, 4, 206
- Hunter, J. D. 2007, *Computing in Science and Engineering*, 9, 90
- Jacob, W. S. 1855, *MNRAS*, 15, 228
- Jara-Maldonado, M., Alarcon-Aquino, V., Rosas-Romero, R., Starostenko, O., & Ramirez-Cortes, J. M. 2020, *Earth Science Informatics*, 13, 573
- Kingma, D. P., & Ba, J. 2014, arXiv e-prints, arXiv:1412.6980
- Kluyver, T., Ragan-Kelley, B., Pérez, F., et al. 2016, in *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, ed. F. Loizides & B. Scmidt (Netherlands: IOS Press), 87–90
- Koch, D. G., Borucki, W. J., Basri, G., et al. 2010, *ApJL*, 713, L79
- Kovács, G., Zucker, S., & Mazeh, T. 2002, *A&A*, 391, 369
- Kunimoto, M., Huang, C., Tey, E., et al. 2021, *Research Notes of the American Astronomical Society*, 5, 234
- Latham, D. W., Mazeh, T., Stefanik, R. P., Mayor, M., & Burki, G. 1989, *Nature*, 339, 38
- Mayor, M., & Queloz, D. 1995, *Nature*, 378, 355
- McCaullif, S. D., Jenkins, J. M., Catanzarite, J., et al. 2015, *ApJ*, 806, 6
- Ofman, L., Averbuch, A., Shliselberg, A., et al. 2022, *NewA*, 91, 101693
- Oliphant, T. E. 2006, *A guide to NumPy*
- Osborn, H. P., Ansdell, M., Ioannou, Y., et al. 2020, *A&A*, 633, A53
- Paegert, M., Stassun, K. G., Collins, K. A., et al. 2021, arXiv e-prints, arXiv:2108.04778
- pandas development team. T. 2020, *pandas-dev/pandas: Pandas*, doi:10.5281/zenodo.3509134
- Pearson, K. A., Palafox, L., & Griffith, C. A. 2018, *MNRAS*, 474, 478
- Pont, F., Zucker, S., & Queloz, D. 2006, *MNRAS*, 373, 231
- Price-Whelan, A. M., Sipőcz, B. M., Günther, H. M., et al. 2018, *AJ*, 156, 123
- Rao, S., Mahabal, A., Rao, N., & Raghavendra, C. 2021, *MNRAS*, 502, 2845
- Ricker, G. R., Winn, J. N., Vanderspek, R., et al. 2015, *Journal of Astronomical Telescopes, Instruments, and Systems*, 1, 014003
- Schanche, N., Collier Cameron, A., Hébrard, G., et al. 2019, *MNRAS*, 483, 5534
- Schwarz, G. 1978, *Annals of Statistics*, 6, 461
- Seabold, S., & Perktold, J. 2010, in *9th Python in Science Conference*
- Shallue, C. J., Lee, J., Antognini, J., et al. 2019, *Journal of Machine Learning Research*, 20, 1
- Shallue, C. J., & Vanderburg, A. 2018, *AJ*, 155, 94
- Song, X., Perel, S., Lee, C., Kochanski, G., & Golovin, D. 2022, in *Automated Machine Learning Conference, Systems Track (AutoML-Conf Systems)*
- Stassun, K. G., Oelkers, R. J., Pepper, J., et al. 2018, *AJ*, 156, 102
- Stassun, K. G., Oelkers, R. J., Paegert, M., et al. 2019, *AJ*, 158, 138

- Tey, E., Moldovan, D., Kunimoto, M., et al. 2022,
Astronet-Triage-v2 dataset, doi:10.5281/zenodo.7411579
- Thompson, S. E., Coughlin, J. L., Hoffman, K., et al. 2018,
ApJS, 235, 38
- Valizadegan, H., Martinho, M., Wilkens, L. S., et al. 2021,
arXiv e-prints, arXiv:2111.10009
- van de Kamp, P. 1963, *AJ*, 68, 515
- Vanderburg, A., & Johnson, J. A. 2014, *PASP*, 126, 948
- Wes McKinney. 2010, in Proceedings of the 9th Python in
Science Conference, ed. Stéfan van der Walt & Jarrod
Millman, 56 – 61
- Wolszczan, A., & Frail, D. A. 1992, *Nature*, 355, 145
- Yu, L., Vanderburg, A., Huang, C., et al. 2019, *AJ*, 158, 25
- Zucker, S., & Giryes, R. 2018, *AJ*, 155, 147

APPENDIX

A. EXAMPLE TCE TABLE

Example TCE table that is passed into `Astronet-Triage-v2` along-side raw light curve data. All data is available in [Tey et al. \(2022\)](#). This table contains information about the signal detected from BLS (epoch, period, duration, depth), information about the host star from TIC 8.2 (TIC ID, M_* , R_* , TMag). Est R_* is described in Section 3.2, and year describes the year the TCE was detected. MinT and MaxT specify the time range used from the light curve for both detection and input to `Astronet-Triage-v2`, and Split specifies which dataset (train, val, test) the signal was in. L1-L8 are labels assigned by individuals and Consensus Label is the label agreed upon by the group.

