The Astronomical Journal, 170:39 (10pp), 2025 July https://doi.org/10.3847/1538-3881/add46d


# Accelerating Long-period Exoplanet Discovery by Combining Deep Learning and Citizen Science

Shreshth A. Malik[1], Nora L. Eisner[2], Ian R. Mason[3], Sofia Platymesi[4], Suzanne Aigrain[4], Stephen J. Roberts[5], Yarin Gal[1], and Chris J. Lintott[4]
[1] OATML, University of Oxford, 7 Parks Road, Oxford, OX1 3QG, UK; shreshth@robots.ox.ac.uk
[2] Center for Computational Astrophysics, Flatiron Institute, 162 Fifth Avenue, New York, NY 10010, USA
[3] Planet Hunters TESS Citizen Scientist, University of Oxford, Keble Road, Oxford, OX1 3RH, UK
[4] Department of Physics, University of Oxford, Keble Road, Oxford, OX1 3RH, UK
[5] Machine Learning Research Group, University of Oxford, Eagle House, Walton Well Road, Oxford, OX2 6ED, UK
*Received 2024 September 20; revised 2025 April 2; accepted 2025 April 17; published 2025 June 19*

## Abstract

Automated planetary transit detection has become vital to identify and prioritize candidates for expert analysis and verification given the scale of modern telescopic surveys. Current methods for short-period exoplanet detection work effectively due to periodicity in the transit signals, but a robust approach for detecting single-transit events is lacking. However, volunteer-labeled transits collected by the Planet Hunters TESS (PHT) project now provide an unprecedented opportunity to investigate a data-driven approach to long-period exoplanet detection. In this work, we train a 1D convolutional neural network to classify planetary transits using PHT volunteer scores as training data. We find that this model recovers planet candidates (TESS objects of interest; TOIs) at a precision and recall rate exceeding those of volunteers, with a 20% improvement in the area under the precision-recall curve and 10% more TOIs identified in the top 500 predictions on average per sector. Importantly, the model also recovers almost all planet candidates found by volunteers but missed by current automated methods (PHT community TOIs). Finally we retrospectively utilise the model to simulate live deployment in PHT to reprioritize candidates for analysis. We also find that multiple promising planet candidates, originally missed by PHT, would have been found using our approach, showing promise for upcoming real-world deployment.

*Unified Astronomy Thesaurus concepts:* Exoplanet detection methods (489); Transit photometry (1709); Convolutional neural networks (1938); Classification (1907); Exoplanet astronomy (486)

## 1. Introduction

Astronomical data sets from recent photometric survey missions, such as the Transiting Exoplanet Survey Satellite (TESS; G. R. Ricker et al. 2014), have grown too large for manual inspection by experts. Detecting planets using the transit method, which involves observing a planet's passage in front of its host star (J. N. Winn 2010; H. J. Deeg & R. Alonso 2018), now typically requires automated analysis. However, most existing algorithms that flag potential exoplanets focus on identifying periodic signals in the star's brightness (G. Kovács et al. 2002), or light curve (LC). This, along with the intrinsic bias of the transit geometry (D. M. Kipping & E. Sandford 2016), leads to an overall bias toward short-period planets and an underrepresentation of longer-period planets in current catalogs.

In response, citizen science projects, that visually inspect large amounts of photometric data in search for transit events, have proven successful in finding planet candidates that have been missed by automated detection methods (D. A. Fischer et al. 2012; M. E. Schwamb et al. 2012; N. L. Eisner et al. 2021; S. M. O'Brien et al. 2024). As the success of detecting a planet candidate via visual inspection is independent of the number of transit events seen in the data, this approach can be used to detect single-transit, longer-period planets. In this work, we investigate how citizen science and machine learning can be combined to improve the recovery of long-period planet candidates, using the Planet Hunters TESS[6] (PHT) citizen science project.

For a complete overview of how planet candidates are identified using PHT, we refer the reader to N. L. Eisner et al. (2021). In brief, PHT is an online citizen science project that has engaged over 45,000 volunteers in the task of visually inspecting the 2 minute cadence TESS light curves in search for transit-like events. Each TESS light curve is seen by 15 volunteers, who independently flag any transit-like signals by drawing a box over them using their mouse. These markings are combined using a density based clustering algorithm, which allows us to identify signals that multiple volunteers flagged. In addition to showing the volunteers real TESS data, they are also randomly shown simulated data—which are simply real TESS light curves with injected simulated transit-like events. These simulated data allow us to assess the skill of each individual volunteer.

By taking into consideration the number of volunteers who flagged any given signal and the skills of these volunteers, we calculate a score for each light curve (between 0 and 1) where a higher score corresponds to a higher confidence in there being a transit-like signal in the given LC. This ranked list allows the PHT science team to prioritize which of the TESS LCs should be evaluated further, where typically the top 500 highest ranked targets are further inspected by experts and grouped into "potential planet candidates," "eclipsing binary" (EB) and "other."

[6] http://www.planethunters.org

This methodology has led to several notable astronomical discoveries (N. L. Eisner et al. 2021, 2022); in particular, longer-period exoplanets identified through single transits. Examples include TOI-2295, a warm Jupiter with a 30 days period (N. Heidari et al. 2025); HD 21520 b, a warm sub-Neptune in a 25 days period (M. Nies et al. 2024); and TOI 4633c, a mini-Neptune in a binary system's habitable zone with a 272 days period (N. L. Eisner et al. 2024).

Aside from citizen science efforts, the most common approach to automated detection is to search for periodic transit signals in the LC (J. M. Jenkins et al. 1996; G. Kovács et al. 2002; A. Collier Cameron et al. 2006; C. Régulo et al. 2007; M. Hippke & R. Heller 2019). The flagged LCs are then validated using a combination of diagnostics, human vetting, and probabilistic and machine learning approaches (B. T. Montet et al. 2015; S. E. Thompson et al. 2015; M. Ansdell et al. 2018; C. J. Shallue & A. Vanderburg 2018; S. Zucker & R. Giryes 2018; L. Yu et al. 2019; H. P. Osborn et al. 2020; D. J. Armstrong et al. 2021; H. Valizadegan et al. 2022).

Machine learning (ML) is a computational approach that enables algorithms to learn statistical regularities directly from data, without being explicitly programmed (T. M. Mitchell & T. M. Mitchell 1997; C. M. Bishop 2007; M. I. Jordan & T. M. Mitchell 2015). In the context of astronomical data analysis, ML algorithms can identify complex patterns and make inferences about new, unseen data by training on existing data sets (N. M. Ball & R. J. Brunner 2010).

Deep learning, a subset of ML, utilizes artificial neural networks with multiple layers to automatically extract and learn rich hierarchical representations of data (Y. LeCun et al. 2015; I. Goodfellow et al. 2016). This is particularly useful in fields with complex and high-dimensional data sets where manual feature extraction methods may not be able to effectively separate signal from noise. Building on the success of deep learning for image (Y. LeCun et al. 1998; A. Krizhevsky et al. 2012; K. Simonyan & A. Zisserman 2015) and audio classification (A. Van Den Oord et al. 2016; H. Ismail Fawaz et al. 2019; S. Kiranyaz et al. 2021), researchers have explored leveraging similar data-driven techniques for exoplanet discovery.

Works have recently sought to go beyond validation and using deep learning directly on the (non-phase-folded) LCs to discover new candidates missed by detection algorithms (K. Cui et al. 2021; G. Olmschenk et al. 2021; M. T. Hansen & J. A. Dittmann 2024). However, the comparatively small number of positive examples has limited the applicability of deep learning for exoplanet discovery (T. A. Hinners et al. 2018). Prior work trains on simulated (synthetic) data to overcome this bottleneck, but this has been shown to have limited applicability on real LCs due to the complex noise processes involved (S. Zucker & R. Giryes 2018). K. Cui et al. (2021) present a promising approach of using 2D object detection algorithms on images of the LC, but are again limited in training data to transits found by automated algorithms. Approaches that leverage different data sources, such as onboard diagnostics (M. T. Hansen & J. A. Dittmann 2024) or applying existing techniques to previously unsearched full-frame images (G. Olmschenk et al. 2021), have also proven effective in identifying transits that were missed by current automated pipelines. Nonetheless, the authors note that they still find a similar distribution of planet candidates to those found by automated algorithms because the training data is biased toward multitransit events.

In this work, we interpret volunteer scores from PHT as transit probabilities, which we use as soft labels to train a 1D convolutional neural network (CNN) classifier to detect transit events from TESS LCs. We find that training with volunteer scores as the main training signal enables the recovery of known planet candidates, with a precision and recall (defined in Section 3.1) significantly better than the volunteers and training with synthetic data. Moreover, the model is able to detect planet candidates missed by traditional automated algorithms and even some that are missed by volunteers.

Our methodology focuses on comprehensive scanning of all available LCs as they are released, generating a prioritized shortlist of candidates for subsequent expert review—a process currently managed exclusively by citizen scientists in the PHT pipeline. Our work provides validation for a future human-in-the-loop machine learning pipeline in exoplanet discovery (B. Settles 2012; M. Walmsley et al. 2020). This approach complements recent work by V. T. Poleo et al. (2024), which further refines such shortlists by further filtering out false positives such as eclipsing binaries. Such refinement is presently conducted by the PHT science team and select citizen scientists before final scientific evaluation.

## 2. Methods

First we describe the data structure of the TESS light curves and PHT labels, and the processing and augmentation steps used in this work (Section 2.1). Then we describe our modeling approach—a 1D CNN architecture we name PlaNet (Section 2.2).

### 2.1. Data

An overview of the data processing pipeline is as follows. (1) Publicly available LC files from the TESS pipeline were downloaded. (2) The LCs were preprocessed for input to the neural network. (3) The LCs were probabilistically transformed using a series of augmentations during training. (4) A proportion of training data was synthetically generated to explicitly include planetary transits and EB false positives. (5) Aggregated volunteer confidence scores for the likelihood of a planetary transit in each LC were collected from the PHT platform to use as ground-truth targets to train the network.

In this work we consider TESS data from sectors 10 to 65. Refer to Appendix A for additional details.

#### 2.1.1. TESS Light Curves

The TESS mission observes around 20,000 stars at a 2 minutes cadence from a new sector of the sky every month (12,000 from Sector 55 onwards). The 2 minute cadence LCs from the Science Processing Operations Center (SPOC) pipeline (J. M. Jenkins et al. 2016) were downloaded from TESS archive (Team M. 2021). We use the Pre-search Data Conditioned Simple Aperture Photometry (PDCSAP) fluxes, which are nominally corrected for instrument variations and flux contamination from nearby stars (D. A. Caldwell et al. 2020). These are thus the best estimate for the intrinsic variation of the star.

*Pre-processing*. The LCs were preprocessed in three steps. (1) Anomaly removal. We used the PDCSAP fluxes (D. A. Caldwell et al. 2020), which are nominally corrected for instrument variations and flux contamination from nearby stars. We also filtered using the `QUALITY` marker given in LC

**Table 1**
Performance of Models Trained on Varying Amounts of Synthetic (Synth.) Data and PHT Volunteer (Vol.) Scores, with and Without the Top 500 Ground-truth (GT) Labels

| Model/Training Data | TOIs | | | | PHT cTOIs | | | | Synthetic Transits versus EBs | |
|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | PR-AUC | R@500 | P@500 | AUC | PR-AUC | R@500 | P@500 | AUC | PR-AUC |
| Volunteers | 0.821(4) | 0.412(8) | 0.52(5) | 0.44(9) | 0.65(3) | 0.0048(11) | 0.32(15) | 0.008(4) | ⋯ | ⋯ |
| **PlaNet (Vol. Scores)** | **0.884(3)** | **0.615(7)** | **0.63(4)** | **0.53(11)** | **0.69(2)** | 0.0038(7) | 0.33(16) | 0.008(4) | 0.513(3) | 0.508(3) |
| Vol. Scores w/o GT | 0.835(4) | 0.430(8) | 0.51(5) | 0.43(9) | 0.65(3) | 0.0033(6) | 0.32(15) | 0.007(2) | 0.382(3) | 0.404(3) |
| 5% Synth. + Vol. Scores | 0.866(4) | 0.579(7) | **0.63(5)** | **0.53(11)** | 0.606(1) | 0.006(1) | 0.38(15) | 0.009(3) | 0.715(2) | 0.718(3) |
| 10% Synth. + Vol. Scores | 0.865(4) | 0.597(7) | **0.63(5)** | **0.53(11)** | 0.57(3) | 0.0046(10) | 0.34(10) | 0.008(3) | 0.711(2) | 0.720(3) |
| 30% Synth. + Vol. Scores | 0.864(4) | 0.499(8) | 0.59(6) | 0.50(10) | 0.55(3) | 0.008(3) | 0.30(11) | 0.007(3) | **0.819(2)** | **0.813(2)** |
| Synth. Only | 0.771(5) | 0.277(7) | 0.38(8) | 0.32(5) | 0.56(3) | **0.009(3)** | 0.32(11) | 0.007(3) | 0.803(2) | 0.804(3) |

**Note.** The performance of the volunteers is also given. The area under the receiver operating characteristic curve (AUC) and the area under the precision-recall curve (PR-AUC) metrics are given for TOIs, PHT community TOIs (cTOIs), and for classifying synthetic transits against synthetic eclipsing binaries (EBs). The uncertainty in the final digit(s) measured as the standard deviation across 1000 bootstrapped samples is given in brackets. The average Recall (R) and Precision (P) at 500 is also given for TOIs and PHT cTOIs, with the uncertainty in the final digit(s) measured as the standard deviation across sectors is given in brackets. There is large variance in PHT cTOIs recovery because of the small number of positives in each sector. In all cases, higher values are better. Statistically significant (t-test with $p < 0.01$) top results are highlighted in bold.

files, which indicate an anomalous event occurred during the measurement such as spacecraft motion or cosmic rays. (2) Binning. For computational reasons we binned the data to a 14 minutes cadence. This is the same cadence as the LCs shown to the volunteers (N. L. Eisner et al. 2021). The duration of transits of interest are generally on the order of hours to days, so the characteristic shape of the dip is still identifiable at this resolution. Empirical tests confirmed that there was no significant classification performance difference in using LCs binned to 6 minutes or 14 minutes. (3) Normalization. We divided and subtracted by the median such that the LC is centered at zero. Dividing by the median (rather than the standard deviation) is common practice in LC analysis as this allows a comparison of the magnitude of brightness dips, which can differentiate between false positives and planet-like transits. We truncated from the start and end (in random proportions) such that all LCs have a binned length of 2500 points (i.e., 24.3 days), and imputed missing values with zeros.

*Augmentations*. During training, we added transformations chosen to help with generalization, with a focus on single transits, each with a probability of 0.1. To simulate a noise process, we randomly chose another LC with volunteer score of 0.0 (so that it is highly unlikely there is a detectable transit present) to inject into the base LC. Three types of data shifts were also incorporated. (1) Two randomly chosen nonoverlapping sections (each 25% of the LC) were switched. (2) The LC was reversed temporally. (3) A random section (10%) of the LC was deleted. These augmentations are particularly helpful to enable generalization across sectors as each sector has systematic noise caused by processes such as momentum dumps that affect all light curves.

*Synthetic Data*. In this work we investigate the efficacy of using volunteer scores as a training signal for single-transit detection. We compare model performance when trained with varying amounts of synthetic data as a comparison—a common approach in prior work (S. Zucker & R. Giryes 2018; K. Cui et al. 2021).

We used synthetic data, which were generated by randomly injecting simulated signals from the ETE-6 data set provided by the SPOC pipeline (J. M. Jenkins et al. 2018) into real TESS light curves. The simulated signals consist of simulated planetary transits and eclipsing binaries (the most common false positive). The same approach is used in the PHT workflow to assess the skill of volunteers (N. L. Eisner et al. 2021).

As we focus on single transits, we only take one transit from each synthetic LC to inject into a random section of the base LC. We used the full flux for EBs as asymmetric dips are often used to identify them.

On injection, we randomly select a transit or EB from the relevant data split and multiply it with the base LC flux at each point before normalization. We use base LCs that have a volunteer score of 0.0 such that they are unlikely to already contain a transit, but we note that these may contain EBs. Unlike in PHT and other work (N. L. Eisner et al. 2021; K. Cui et al. 2021), we do not place a minimum signal to noise (SNR) constraint on injected transits as we hope to be able to identify shallow transits. We do, however, place a maximum SNR constraint of 15, and a maximum duration of 4 days (as in PHT). Refer to Appendix A.2 for additional details.

#### 2.1.2. Volunteer Scores and Planetary Candidate Labels

We used the aggregated confidence scores from PHT that take into account the volunteer skill level (N. L. Eisner et al. 2021) as soft targets (between 0 and 1) to train the network for binary classification. The 500 LCs with the highest volunteer scores for each sector are then analyzed by experts, who classify them as high-potential planet candidates for in-depth vetting, EBs, or false positives, as described in Section 1. We use these "ground-truth" labels to correct for highly confident, but incorrect, volunteer scores. This correction significantly increased performance across candidates (Table 1).

We also cross-referenced the star corresponding to each LC with discrete classifications of planetary candidates from TESS for evaluation. TESS observes a new patch (a sector) of the sky every month and sends the sensor data to Earth for post-processing. An automated pipeline (J. M. Jenkins et al. 2016) extracts and processes the raw sensor readings into 1D LCs for each star in the sector. This pipeline flags ∼7% of the observations as threshold crossing events (TCEs), which

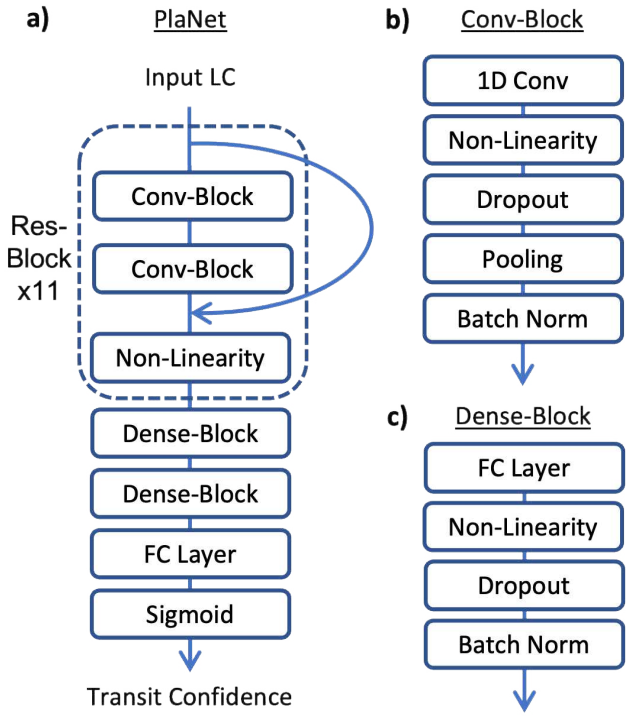


**Figure 1.** (a) Simplified schematic of the proposed PlaNet architecture featuring a series of Convolutional Blocks (Conv-Blocks) with residual connections (K. He et al. 2016), followed by two Dense-Blocks. (b) Ordering of layers in the Conv-Blocks. (c) Ordering of layers in the Dense-Blocks. See main text for details and Appendix B for hyperparameters.

indicates the presence of a periodic signal. TCEs are then analyzed by the TESS team and ∼18% are promoted to TESS Objects of Interest (TOIs) status as likely planetary candidates. Similarly, candidates that have been identified through volunteer flagging and vetted by the PHT science team are called Community TOIs (PHT cTOIs). Overall, ∼1% of all LCs have planetary transits, whereas ∼21% have a nonzero volunteer confidence scores (Appendix A). Thus the soft labels provide more training signal to the model.

### 2.2. Model

We now describe our PlaNet model and training procedure. Appendix B contains specific implementation details.

*Architecture*. PlaNet is a one-dimensional CNN designed to classify LCs by determining the likelihood of a planetary transit event. The architecture (Figure 1), consists of 11 residual blocks, each consisting of two convolutional neural network blocks (Y. LeCun et al. 2015; J. Schmidhuber 2015) with a skip connection (K. He et al. 2016). These "residual" connections add the input to the block back to the output of the block, which helps mitigate the vanishing gradient problem that can hinder training in deep networks, where gradient information diminishes as the gradient is propagated backward through the network (K. He et al. 2016). These are followed by two dense blocks and finally a fully-connected layer with sigmoid activation to output a transit confidence score between 0 and 1.

In comparison to prior work (G. Olmschenk et al. 2021), we used a deeper network (22 convolutional layers) with residual connections and larger kernel sizes in the first two layers (7 and 5, respectively) without down sampling, to increase receptive field. The receptive field refers to the range of input data that a neuron can process, and a larger receptive field allows the model to capture more extensive temporal dependencies in the data. These modifications resulted in significant performance gains over the architecture in G. Olmschenk et al. (2021).

Appendix B contains further details on the model. Our code is publicly available.[7]

*Training and Validation*. To evaluate the model's performance in a realistic deployment scenario, we partitioned the data by TESS observational sectors: sectors 10–44 for training, sectors 45–54 for validation, and sectors 55–65 for testing. This division ensures that the model is tested on data from sectors it has not encountered during training, simulating its application to future observations. We also randomly split the available synthetic transits and EBs to ensure no test-set leakage occurs. A detailed breakdown of the data split is provided in Table 3 in Appendix A.

The model was trained using gradient descent with the binary cross-entropy loss function, a standard choice for binary classification tasks (C. M. Bishop 2007). It is defined as $L = -(y \log \hat{y} + (1 - y)\log(1 - \hat{y}))$ for a single example, where $y$ is the true label and $\hat{y}$ is the model's predicted probability. In our case, the true labels are "soft" volunteer labels ranging between 0 and 1, indicating the confidence level of a transit event in each light curve. Minimizing this loss encourages the network to predict transit probabilities that align with the volunteer distribution.

We conducted a heuristic search over various architectural and optimization hyperparameters, selecting the model that achieved the lowest validation loss during training through early stopping. Early stopping involves monitoring the model's performance on the validation set and terminating training when performance ceases to improve. This prevents overfitting to the training set, and therefore improves generalization (I. Goodfellow et al. 2016).

## 3. Results

In this section we discuss the performance of the proposed model. We compare its ability to recover known candidate planets (TOIs) to volunteers (Section 3.2). We then analyze how using volunteer scores for training the model compares to using synthetic data, as is commonplace in existing work (Section 3.3). Next, we show that using the model in the analysis pipeline would have led to the recovery of additional planet candidates that were originally missed by the PHT pipeline (Section 3.4). Finally, we qualitatively compare volunteer and PlaNet predictions (Section 3.5).

### 3.1. Metrics

To measure model performance quantitatively, we use the following metrics throughout the paper:

1. The fractional recovery rate or recall for the top K predictions (R@K) is the proportion of the class (e.g., TOIs) that are recovered in the top-ranked predictions, defined as the number of true positives divided by the total number of positive examples in the full data set. A

[7] 10.5281/zenodo.4311816

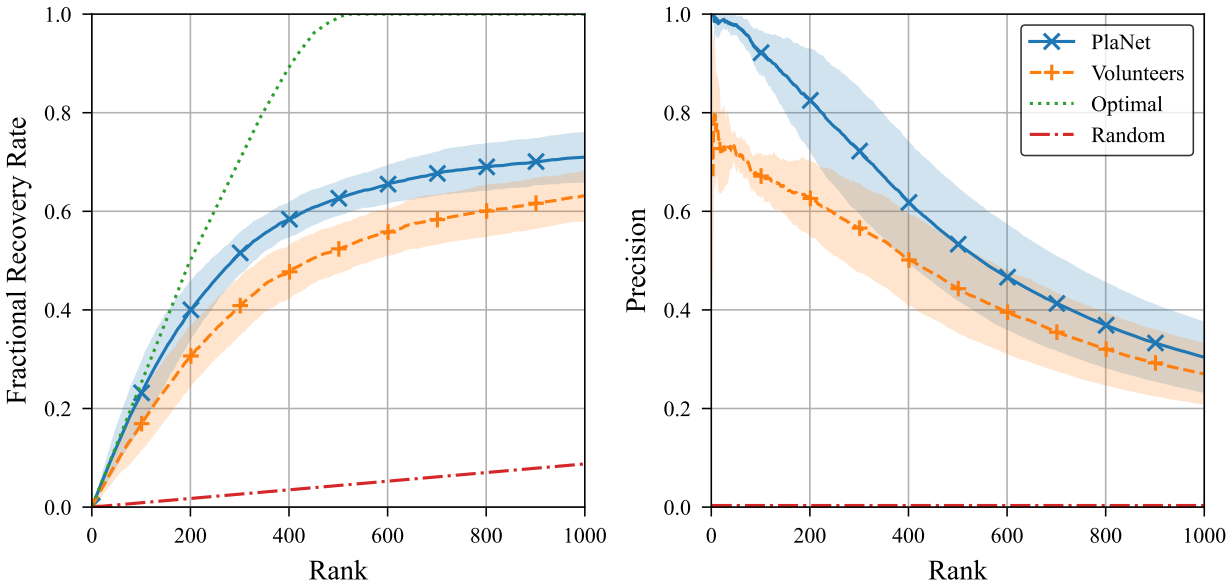


**Figure 2.** Fractional recovery of TOIs (left) and precision (right) of top-ranked predictions, averaged across test sectors. The shaded areas indicate the standard deviation across sectors. The optimal and random curves are averages across sectors. PlaNet significantly outperforms Volunteers in both precision and recall.

high R@K indicates that we recover most of the candidates within the top K predictions.

2. The precision for the top K predictions (P@K) is a measure of model accuracy and is defined as the number of true positives divided by the number of positive predictions (true positives and false positives). A high P@K indicates that few of the top-ranked predictions are false positives.
3. The area under the receiver operator curve (AUC; J. A. Hanley & B. J. McNeil 1982) is a summary scalar metric for the performance of a binary classifier. It is computed by calculating the true positive rate and false positive rate at various decision thresholds, and then computing the area underneath this curve. An AUC of 1 would indicate a perfect classifier, whereas an AUC of 0.5 would indicate a random classifier.
4. The area under the precision-recall curve (PR-AUC) is defined similarly to the AUC. The precision and recall are calculated for different decision thresholds and the area under this curve is computed.

The P@K and R@K metrics are particularly important in our context as our aim is to use the classifier to rank LCs by their score and shortlist the top predictions. AUC and PR-AUC are useful to measure overall skill of the classifier, balancing both precision and recall.

### 3.2. Known Candidate Recovery

First we consider the recovery of TOIs (known planet candidates) in the test sectors. Figure 2 show the fractional recovery rate of TOIs and the precision of predictions for the top 1000 (∼5%) ranked predictions by the model compared to that of the volunteers. We see that the model recovers TOIs at a rate and precision significantly better than volunteers.

To put this into context, in PHT, the top 500 ranked LCs in each sector are shortlisted for further analysis. If PlaNet was used instead of the volunteer scores, this would have resulted in ∼10% more TOIs recovered on average, with a ∼5% improvement in precision. Only one of the top ∼50 candidates in each sector is not a TOI on average (precision of 0.98), compared to volunteers which have >1 in 4 false positives (precision of 0.72).

Figure 3 and Table 1 show classification performance of the model compared to volunteers on TOIs and PHT cTOIs. We see that the model also recovers PHT cTOIs at a rate similar to, or exceeding volunteers. This is important as we hope that this model can augment the PHT project to recover planet candidates outside of TOI distribution. Table 1 also shows that correcting for highly confident but incorrect volunteer labels significantly boosts model performance.

### 3.3. Comparison to Synthetic Data Training

Table 1 shows the performance of models on both real and synthetic data. For evaluation on synthetic data, we compare performance at identifying transits from a balanced test data set of synthetic EBs and synthetic transits.

We find that using volunteer scores training data enables significantly higher performance on TOI recovery than training with synthetic data only. As our synthetic data is focused on single transits, we find that PHT cTOIs recall and precision is similar across models. However as there is only a small number of PHT cTOIs per sector (<5), this is not enough to determine which model is better at recovery. Nonetheless, the PHT cTOIs AUC decreases monotonically as the proportion of synthetic data is increased, which indicates that the model performance degrades on this class when more synthetic data is used.

We note that the performance on synthetic data (classifying synthetic transits against synthetic EBs) of a model trained only on volunteer scores is significantly worse. Including some synthetic data boost results on synthetic data significantly. This is unsurprising; training on synthetic data improves performance on synthetic data. However, if the goal is to recover real planets, optimizing this metric can therefore be deceiving.

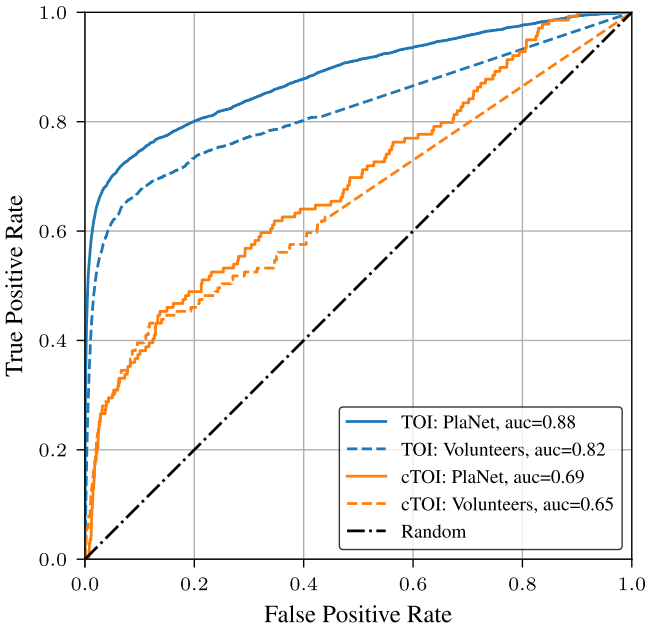


**Figure 3.** Receiver operating characteristic for TOIs and PHT cTOIs. PlaNet outperforms Volunteers on both classes.

### *3.4. Identifying Previously Missed Candidates*

In the previous sections we have shown that training on volunteer labels is more effective than synthetic data at recovering TOIs. However, our primary objective is to find novel candidates. In this section we evaluate this by retrospectively analyzing the top model predictions in each of the test sectors, as it would be in live deployment.

We took the top 500 most confident predictions for each sector once known TOIs have been removed (as in PHT). We found that almost all (19 out of 23) PHT cTOIs identified by volunteers remained in this top 500 per sector. Notably, these are single-transit events which are missed by existing algorithms. By combining model and volunteer scores (further discussion in Section 4), it is hoped that no planet candidates will be missed in the final pipeline.

To find if the model had recovered any *additional* candidates that were missed by the PHT pipeline, we then removed any LCs which also appeared in the volunteer labeled top 500 which had already been analyzed. This left approximately 200 per sector. We manually analyzed the top 30 remaining predictions per sector to see whether there were planet candidates that were missed by the initial PHT pipeline but would have been found by the new pipeline.

From this investigation, we found three planet candidates that were previously missed by the PHT pipeline, shown in Figure 4 and Table 2. For each of these candidates, we performed a number of standard diagnostic tests to help rule out instrumental and astrophysical false positive scenarios including background eclipsing binaries, systematic effects, and background events such as asteroids passing through the field of view (for details, see N. Eisner et al. 2020). All of the candidates passed these tests. We do note that these candidates were also identified by other teams: the Warm gIaNts with TESS (WINE) collaboration (M. J. Hobson et al. 2021) and the TESS Single Transit Planet Candidates (TSTPC) working group (S. Villanueva et al. 2019; I. Mireles et al. 2023), and have already been uploaded to ExoFOP (ExoFOP 2019) as

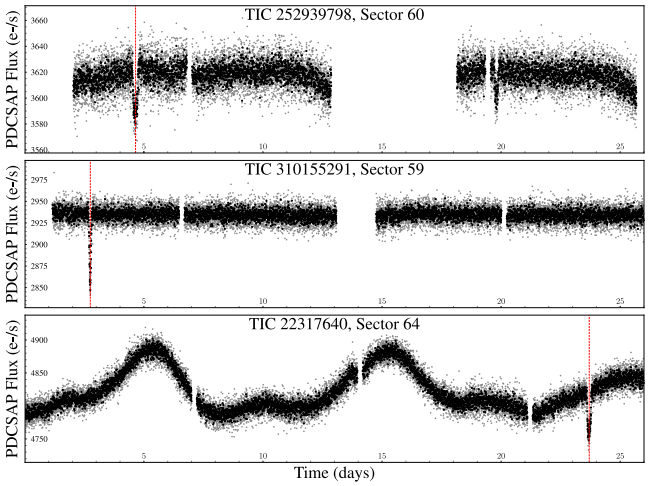


**Figure 4.** Candidates (cTOIs) found by PlaNet, originally found by other teams. See Table 4 for details.

cTOIs. Nonetheless, this analysis shows that PlaNet is able to help identify new planet candidates that are missed by the standard PHT pipeline. We hypothesize that TIC 252939798 and 310155291 were not originally found by the PHT pipeline because they could be mistaken for EBs. We note however, the high volunteer score for TIC 22317640. This was likely originally dismissed as it was found by a previous team already.

Given the improvement in TOI recovery discussed in Section 3.2, one might expect to find more missed candidates in the test sectors. However, this is largely constrained by the small number of PHT cTOIs overall and the fact that only 30 additional light curves per sector were analyzed. It is also worth noting that most of the false positives among the top-ranked results are EBs. Further filtering using additional metadata could help eliminate more EBs, potentially increasing the yield of planet candidates (V. T. Poleo et al. 2024).

### *3.5. Qualitative Analysis*

In Figure 5 we present examples of volunteer and model predictions on test sectors. We find that the models recover single-transit planet candidates that the volunteers find (Figure 5(b)) as well as those which were missed (Figure 5(a)).

On inspection of the PHT cTOIs that were missed by PlaNet (i.e., not in the top 500 predictions after removal of TOIs), for example, Figure 5(c), we find that these are predominantly shallow single-transit events. Up-weighting losses for shallow transits could improve results.

The majority of false positives (Figure 5(d)) from both PlaNet and volunteers tend to be less-obvious eclipsing binaries. These type of false positives can be prevented by adding further features to the model such as background flux and stellar metadata to better vet the most promising candidates (V. T. Poleo et al. 2024).

## 4. Discussion

In Section 3.3, we demonstrated that training models using synthetic data, as employed in PHT and other studies, yields lower performance compared to training with volunteer labels. Despite the inherent noise in the volunteer labels, machine learning models trained on this data outperform those trained on synthetic data from the ETE-6 data set in recovering planet

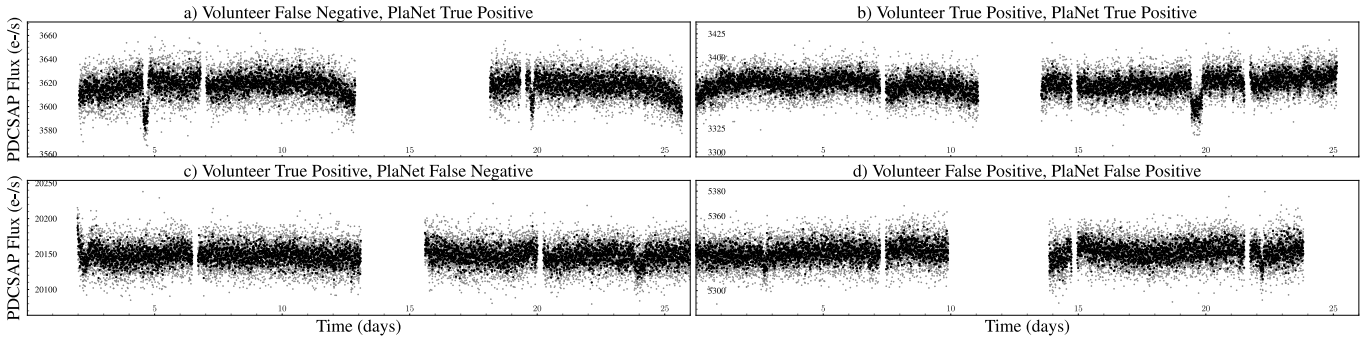


**Figure 5.** Volunteer and PlaNet model predictions on example light curves in test sectors. (a) TIC 252939798, sector 60. An example of a cTOI found by PlaNet but not by the volunteer pipeline. (b) TIC 457351794 Sector 65. An example of a single-transit PHT cTOI found by both volunteers and the model. (c) TIC 284361752 Sector 59. An example of a PHT cTOI missed by PlaNet but found by the volunteer pipeline. This is a shallow transit event. (d) TIC 150031040, Sector 65. An example of a false positive (eclipsing binary) for both volunteers and the model.

**Table 2**
Details of cTOIs Found by PlaNet, Originally Found by Other Teams

| TIC ID | PHT Score | Depth (mmag) | Depth (ppm) | Radius ($R_\oplus$) | Pipeline/Team |
|---|---|---|---|---|---|
| 252939798 | 0.88 | 8.52 ± 1.29 | 7813 ± 1192 | 10.12 ± 0.93 | TSTPC Working Group (S. Villanueva et al. 2019) |
| 310155291 | 0.57 | 40.16 ± 21.39 | 36312 ± 19511 | 15.16 ± 4.14 | TSTPC Working Group (S. Villanueva et al. 2019) |
| 22317640 | 0.94 | 7.41 ± 0.33 | 6800 ± 300 | ⋯ | WINE collaboration (M. J. Hobson et al. 2021) |

**Note.** Light curves are shown in Figure 4. All information was retrieved from ExoFOP (ExoFOP 2019).

candidates. This highlights the need for higher quality synthetic data when the ultimate objective is to detect real planets. Synthetic data quality can increase, for example, by using taking advantage of advances in deep generative modeling (J. Sohl-Dickstein et al. 2015; J. Ho et al. 2020; J. Song et al. 2021). These models could be used to generate synthetic LCs which are more realistic than ETE-6, thereby enhancing the training of classifiers for real planet detection. Alternatively, greater emphasis could be placed on training with more informative signals, such as citizen science labels.

Once PlaNet has been trained, generating predictions on all the LCs in a sector takes minutes on a single GPU. Hence inference can be computed as soon as new sector LC data is available from TESS to shortlist candidates for further vetting before waiting for volunteers to label. This can reduce the time from data availability to planet candidate identification drastically.

Improving the way we used the citizen science labels could significantly improve performance. Currently we use the aggregated PHT score as a single label per LC. However, we have a distribution of scores from each volunteer per light curve. Incorporating this uncertainty into the labels could improve performance. Moreover, we could incorporate auxiliary confidence losses (C. Burns et al. 2024), or a temperature $T$ to the labels $y \rightarrow \exp(y/T)$ to adjust the entropy of the training label distribution, forcing the model to be more confident in its predictions.

Volunteer numbers do not scale at the same rate as the data generated by modern exoplanet surveys. Hence, as we progress to larger surveys (Ž. Ivezić et al. 2019; Y. Wang et al. 2022) it is important to prioritize volunteer time as well. PlaNet can be used as a pre-filtering mechanism for the PHT platform to remove confidently negative LCs from volunteer labeling workflows. Incorporating calibrated uncertainty estimation would help identify costly overconfident predictions (Y. Gal & Z. Ghahramani 2016; B. Lakshminarayanan et al. 2017). This will also enable an active learning pipeline to be developed which can make better use of volunteer time (B. Settles 2012; M. Walmsley et al. 2020).

Furthermore, as discussed in Section 3.4, we found considerable overlap (approximately 60%) between the top 500 ranked predictions per sector made by volunteers and the model. However, the remaining predictions suggest some divergence in what volunteers and the model identify as high-potential transits, as shown qualitatively in Section 3.5. We therefore propose that combining scores from both human and machine assessments (D. E. Wright et al. 2017) could leverage the strengths of each to recover more candidates in future sectors.

The main focus of this work was to access the efficacy of citizen science labels as training data for exoplanet recovery. However, there is much that can be done to also improve the PlaNet model architecturally. We can investigate using wavelet based models (H. Liao et al. 2024), object recognition models (K. Cui et al. 2021), or transformers and self-supervised pretrained models for light curves (C. Donoso-Oliva et al. 2023), similar to those used in audio classification (A. Baevski et al. 2020). We can also add further features such as background flux and stellar metadata to improve predictive ability (D. J. Armstrong et al. 2021; V. T. Poleo et al. 2024). We note our approach of using citizen science labels for training is agnostic to the architecture of the model used.

## 5. Conclusion

In this work, we used volunteer scores from Planet Hunters TESS to train a 1D CNN to detect planetary transits from TESS light curves. The model was found to exceed the original volunteer's recovery rate on TOIs, and is significantly better than one trained on synthetic data. Moreover, as observed in N. L. Eisner et al. (2021), we recover single-transit candidates that are missed by traditional pipelines, and even some that were missed by volunteers.

The model is to be integrated into the PHT pipeline for exoplanet discovery, where it will serve as a basis for a human-in-the-loop ML pipeline to more efficiently shortlist candidates for further vetting by machine learning methods (V. T. Poleo et al. 2024) and finally review by scientists.

## Acknowledgments

S.M. led the project, wrote the code, carried out all the experiments and wrote the paper. N.E. provided the PHT data and exoplanet domain expertize, and along with IM, conducted candidate analysis and vetting. S.A. and S.P. are members of the PHT team. C.L. and Y.G. initiated the project, and along with SR, provided high-level guidance. All authors provided critical feedback on writing.

We like to thank all of the volunteers who participated in the Planet Hunters TESS project. Some/all of the data presented in this paper were obtained from the Mikulski Archive for Space Telescopes (MAST). STScI is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS5-26555. Support for MAST for non-HST data is provided by the NASA Office of Space Science via grant NNX13AC07G and by other grants and contracts. Planet Hunters TESS is supported in part by the Alfred P. Sloan Foundation.

S.M. acknowledges funding from EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines and Systems (grant No: EP/S024050/1). Y.G. is supported by a Turing AI Fellowship financed by the UK government's Office for Artificial Intelligence, through UK Research and Innovation (grant reference EP/V030302/1) and delivered by the Alan Turing Institute.

## Appendix A
## Data

### A.1. Data Composition

Table 3 shows the breakdown of the sector, volunteer score, planet candidate, and synthetic data distributions in each of the data splits. Planets and EBs with TIC IDs multiple of 4 were chosen for synthetic data generation in the training set, and those with a remainder of 1 for the validation set. The remainder were chosen for the test set.

**Table 3**
Breakdown of the Examples in Each of the Data Splits used in this Work

| Split | Sectors | Total LCs | TOIs | PHT cTOIs | Vol. Scores $>$ **0** | Vol. Scores $>$ **0.5** | Synth. Transits | Synth. EBs |
|---|---|---|---|---|---|---|---|---|
| Training | 10–44 | 643,412 | 7170 | 472 | 137,818 | 16,609 | 278 | 65 |
| Validation | 45–54 | 139,976 | 2377 | 116 | 44,080 | 3383 | 316 | 69 |
| Test | 55–65 | 143,487 | 4701 | 139 | 65,791 | 6229 | 616 | 145 |

**Note.** The number of synthetic transits and eclipsing binaries used to generate the synthetic data is also given. The number of (c)TOIs correspond to the number of LCs which are associated with a (c)TOI; some objects are observed multiple times across sectors, so the total number of (c)TOIs in the table is greater than the actual number of (c)TOIs.

### A.2. Synthetic Data

When synthetic data was used, an equal proportion of synthetic transits to synthetic EBs was used in all cases. In the synthetic data composition experiments (Section 3.3), the model that uses only synthetic data uses 30% synthetic transits, 30% EBs. The remaining 40% of LCs were chosen such that they were unlikely to contain a transit to balance the class split. This was done by selecting LCs with a volunteer score of 0.0. The test synthetic data set used for evaluation in Table 1 was created with an equal mix of only injected transits and EBs in the test sector LCs. Note that these are probabilistic synthetic injections, and as we only inject EBs or transits into LCs with a volunteer score of zero, the actual proportion of synthetic data used is lower than the percentage given.

When single-transit synthetics were injected into the base LC, we ensured that at least 80% of the section of the base LC where the transit is being injected into was not missing data. This is to prevent injecting a transit into a missing data region where it would not be visible and thus be mislabeled.

The SNR for synthetic data was calculated as the depth of the injected transit divided by the Combined Differential Photometric Precision (CDPP) of the base light curve. CDPP is the metric that defines the ease with which these weak terrestrial transit signatures can be detected. This is given at a series of durations (0.5, 1, 2 days). The closest one to the duration of the transits is used to calculate the SNR.

## Appendix B
## Implementation Details

Models were implemented in PyTorch (A. Paszke et al. 2019). The main PlaNet model hyperparameters are given in Table 4. Architecture and optimization hyperparameters were chosen heuristically through minimizing validation loss. See Section 2.2 for details of the architecture. Along with the changes described compared to existing work, we also tried experimented with larger kernels and dilated convolutions (A. Van Den Oord et al. 2016). Dilated convolutions involve inserting gaps between the kernel elements, effectively expanding the kernel's receptive field without increasing the number of parameters. However, these modifications did not yield significant improvements in performance. We specifically did not use a fully convolutional model as some degree of temporal awareness is required to identify EB false positives with multiple dips. Training runs on the full data set took approximately 16 hr on a single NVIDIA GeForce GTX 2080 Ti GPU. Our code is open source and available at 10.5281/zenodo.4311816, with ongoing development at https://github.com/s-a-malik/pht-ml.

**Table 4**
Hyperparameter Configuration for PlaNet

| Hyperparameter | Value |
|---|---|
| Optimization | |
| Optimizer | AdamW (I. Loshchilov & F. Hutter 2019) |
| L2 weight regularization | 0.01 |
| Learning rate | 0.001 |
| Batch size | 128 |
| Early stopping patience (epochs) | 100 |
| Dropout rate | 0.1 |
| Max epochs | 300 |
| Architecture | |
| Hidden layer nonlinearities | Leaky-ReLU (slope = 0.01) |
| Layers | [Res-Block (x11), Dense-Block (x2), Linear] |
| Kernel size | [7, 5, 3, 3, 3, 3, 3, 3, 3, 3, 3, N/A, N/A, N/A] |
| Dropout (true/false) | [F, T, T, T, T, T, T, T, T, T, T, F, N/A] |
| Batch normalization (true/false) | [F, T, T, T, T, T, T, T, T, T, T, F, N/A] |
| Number of out channels/units | [32, 32, 32, 64, 64, 128, 128, 128, 128, 128, 128, 1280, 256, 20] |
| Max pooling size | [1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 1, N/A, N/A, N/A] |
| Total number of parameters | 1,054,889 |
| Data | |
| Synthetic transit proportion | 0.0 |
| Synthetic EB proportion | 0.0 |
| Training augmentation probability | 0.1 |
| Bin factor (max LC length) | 7 (2500) |
| Permute fraction in training | 0.25 |
| Delete fraction in training | 0.1 |

## ORCID iDs

Shreshth A. Malik https://orcid.org/0000-0003-1544-3050
Nora L. Eisner https://orcid.org/0000-0002-9138-9028
Ian R. Mason https://orcid.org/0009-0000-4995-8875
Sofia Platymesi https://orcid.org/0009-0007-6871-0008
Suzanne Aigrain https://orcid.org/0000-0003-1453-0574
Stephen J. Roberts https://orcid.org/0000-0002-9305-9268
Yarin Gal https://orcid.org/0000-0002-2733-2078
Chris J. Lintott https://orcid.org/0000-0001-5578-359X

## References

Ansdell, M., Ioannou, Y., Osborn, H. P., et al. 2018, ApJL, 869, L7
Armstrong, D. J., Gamper, J., & Damoulas, T. 2021, MNRAS, 504, 5327
Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. 2020, in Adv. Neural Inf. Process Syst., Vol. 33, ed. H. Larochelle et al. (Red Hook, NY: Curran Associates, Inc.), 12449
Ball, N. M., & Brunner, R. J. 2010, IJMPD, 19, 1049
Bishop, C. M. 2007, Pattern Recognition and Machine Learning (Berlin: Springer)
Burns, C., Izmailov, P., Kirchner, J. H., et al. 2024, in ICML, Vol. 41 (Vienna: JMLR), 196
Caldwell, D. A., Tenenbaum, P., Twicken, J. D., et al. 2020, RNAAS, 4, 201
Collier Cameron, A., Pollacco, D., Street, R. A., et al. 2006, MNRAS, 373, 799
Cui, K., Liu, J., Feng, F., & Liu, J. 2021, AJ, 163, 23
Deeg, H. J., & Alonso, R. 2018, in Transit Photometry as an Exoplanet Discovery Method, ed. H. J. Deeg & J. A. Belmonte (Cham: Springer), 633
Donoso-Oliva, C., Becker, I., Protopapas, P., et al. 2023, A&A, 670, A54
Eisner, N., Lintott, C., & Aigrain, S. 2020, JOSS, 5, 2101
Eisner, N. L., Barragán, O., Lintott, C., et al. 2021, MNRAS, 501, 4669
Eisner, N. L., Johnston, C., Toonen, S., et al. 2022, MNRAS, 511, 4710
Eisner, N. L., Grunblatt, S. K., Barragán, O., et al. 2024, AJ, 167, 241
ExoFOP 2019, Exoplanet Follow-up Observing Program - TESS, IPAC, doi:10.26134/EXOFOP3
Fischer, D. A., Schwamb, M. E., Schawinski, K., et al. 2012, MNRAS, 419, 2900
Gal, Y., & Ghahramani, Z. 2016, in ICML, Vol. 33, ed. M. F. Balcan & K. Q. Weinberger (New York: JMLR), 1050
Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. 2016, Deep Learning (Cambridge, MA: MIT Press)
Hanley, J. A., & McNeil, B. J. 1982, Radiology, 143, 29
Hansen, M. T., & Dittmann, J. A. 2024, AJ, 168, 291
He, K., Zhang, X., Ren, S., & Sun, J. 2016, in Proc. the IEEE Conf. on Computer Vision and Pattern Recognition (New York: IEEE), 770
Heidari, N., Hébrard, G., Martioli, E., et al. 2025, A&A, 694, A36
Hinners, T. A., Tat, K., & Thorp, R. 2018, AJ, 156, 7
Hippke, M., & Heller, R. 2019, A&A, 623, A39
Ho, J., Jain, A., & Abbeel, P. 2020, in Adv. Neural Inf. Process Syst., 33, Virtual ed. H, Larochelle et al. (Red Hook, NY: Curran Associates, Inc.), 6840
Hobson, M. J., Brahm, R., Jordan, A., et al. 2021, in Posters from the TESS Science Conf. II (TSC2) (Genève: Zenodo), 25
Ismail Fawaz, H., Forestier, G., Weber, J., Idoumghar, L., & Muller, P.-A. 2019, Data Min. Knowl., 33, 917
Ivezić, Ž., Kahn, S. M., Tyson, J. A., et al. 2019, ApJ, 873, 111
Jenkins, J. M., Doyle, L. R., & Cullers, D. K. 1996, Icar, 119, 244
Jenkins, J. M., Twicken, J. D., McCauliff, S., et al. 2016, Proc. SPIE, 9913, 1232
Jenkins, J. M., Tenenbaum, P., Caldwell, D. A., et al. 2018, RNAAS, 2, 47
Jordan, M. I., & Mitchell, T. M. 2015, Sci, 349, 255
Kipping, D. M., & Sandford, E. 2016, MNRAS, 463, 1323
Kiranyaz, S., Avci, O., Abdeljaber, O., et al. 2021, MSSP, 151, 107398
Kovács, G., Zucker, S., & Mazeh, T. 2002, A&A, 391, 369
Krizhevsky, A., Sutskever, I., & Hinton, G. E. 2012, in Adv. Neural Inf. Process Syst., Vol. 26, ed. P. L. Bartlett et al. (Red Hook, NY: Curran Associates, Inc.), 1106
Lakshminarayanan, B., Pritzel, A., & Blundell, C. 2017, in Adv. Neural Inf. Process Syst., Vol. 30, ed. I. Guyon et al. (Red Hook, NY: Curran Associates, Inc.), 6402
LeCun, Y., Bengio, Y., & Hinton, G. 2015, Natur, 521, 436
LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. 1998, Proc. IEEE, 86, 2278
Liao, H., Ren, G., Chen, X., Li, Y., & Li, G. 2024, AJ, 167, 180
Loshchilov, I., & Hutter, F. 2019, arXiv:1711.05101
Mireles, I., Dragomir, D., Osborn, H. P., et al. 2023, ApJL, 954, L15
Mitchell, T. M. 1997, Machine learning (New York: McGraw-Hill)
Montet, B. T., Morton, T. D., Foreman-Mackey, D., et al. 2015, ApJ, 809, 25
Nies, M., Mireles, I., Bouchy, F., et al. 2024, MNRAS, 534, 3744
O'Brien, S. M., Schwamb, M. E., Gill, S., et al. 2024, AJ, 167, 238
Olmschenk, G., Silva, S. I., Rau, G., et al. 2021, AJ, 161, 273
Osborn, H. P., Ansdell, M., Ioannou, Y., et al. 2020, A&A, 633, A53
Paszke, A., Gross, S., Massa, F., et al. 2019, in Adv. Neural Inf. Process Syst., Vol. 32, ed. H. Wallach et al. (Red Hook, NY: Curran Associates, Inc.), 8024, https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf
Poleo, V. T., Eisner, N., & Hogg, D. W. 2024, AJ, 168, 100
Régulo, C., Almenara, J. M., Alonso, R., Deeg, H., & Roca Cortés, T. 2007, A&A, 467, 1345
Ricker, G. R., Winn, J. N., Vanderspek, R., et al. 2014, JATIS, 1, 014003
Schmidhuber, J. 2015, NN, 61, 85
Schwamb, M. E., Lintott, C. J., Fischer, D. A., et al. 2012, ApJ, 754, 129
Settles, B. 2012, Active Learning, Vol. 6 (Berlin: Springer)
Shallue, C. J., & Vanderburg, A. 2018, AJ, 155, 94
Simonyan, K., & Zisserman, A. 2015, in ICLR, Vol. 3, ed. Y. Bengio & Y. LeCun (San Diego, CA: OpenReview)
Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. 2015, in ICML, Vol. 32, ed. F. Bach & D. Blei (Lille: JMLR), 2256
Song, J., Meng, C., & Ermon, S. 2021, in ICLR, Vol. 9 (Lille: OpenReview)
Team M. 2021, TESS Light Curves - All Sectors, STScI/MAST, doi:10.17909/T9-NMC8-F686
Thompson, S. E., Mullally, F., Coughlin, J., et al. 2015, ApJ, 812, 46
Valizadegan, H., Martinho, M. J., Wilkens, L. S., et al. 2022, ApJ, 926, 120
Van Den Oord, A., Dieleman, S., Zen, H., et al. 2016, arXiv:1609.03499
Villanueva, S., Jr., Dragomir, D., & Gaudi, B. S. 2019, AJ, 157, 84
Walmsley, M., Smith, L., Lintott, C., et al. 2020, MNRAS, 491, 1554
Wang, Y., Zhai, Z., Alavi, A., et al. 2022, ApJ, 928, 1
Winn, J. N. 2010, arXiv:1001.2010
Wright, D. E., Lintott, C. J., Smartt, S. J., et al. 2017, MNRAS, 472, 1315
Yu, L., Vanderburg, A., Huang, C., et al. 2019, AJ, 158, 25
Zucker, S., & Giryes, R. 2018, AJ, 155, 147