MNRAS **000**, 1–?? (2016) Preprint 22 September 2016 Compiled using MNRAS LATEX style file v3.0

A transient search using combined human and machine classifications

Darryl Wright,[1⋆] Chris Lintott,[1] Ken Smith,[2]
[1]*Department of Physics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford, OX1 3RH* [2]*QUBs*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

Large modern surveys require efficient review of data in order to find transient sources such as supernovae, and to distinguish such sources from artefacts, noise and so on. Much effort has been put into the development of automatic algorithms, but surveys still rely on human review of targets. This paper presents an integrated system for the identification of supernovae in data from PanSTARRS, combining classifications from a citizen science project including volunteers with those from a convolutional neural network. This work represents the first time such a system has been deployed on real astronomical data, and we show that the combination of the two methods outperforms either one used individually. This result has important implications for the future development of transit searches, especially in the era of LSST and other large-throughput surveys.

Key words: keyword1 – keyword2 – keyword3

1 INTRODUCTION

CJL

The detection and identification of transient sources has long been an important part of astronomical observation. New surveys such as LSST (XXXXXX CITE XXXXX) will increase the number of transient candidates detected by many orders of magnitude, leading to renewed attention being paid to the methods used by transient searches. To extract the most scientific value from transients we want to follow their entire evolution from the time of outburst to the point at which they fade below the detection limit. This requires a rapid decision on whether or not to expend valuable followup resources for each potential candidate extracted by a transient survey's image processing pipeline. The first problem is deciding if a source, flagged by the pipeline, is a detection with real astrophysical significance or an artefact of the detector or image processing. We want to promote the former for a decision on whether to follow and to reject the latter without further consideration. In preparation for LSST and to deal with the data volumes of present surveys, much effort has been invested to automatically reject false positives with supervised learning. Using large volumes of past observations that have been ground truthed as real or "bogus", the aim is to train a machine to make predictions about future observations (Bloom et al. 2012; Brink et al. 2013; Goldstein et al. 2015; du Buisson et al. 2015; Donalek et al. 2008; Romano et al. 2006; Bailey et al. 2007).

Large quantities of labelled data can be time consuming to generate for small science teams. An alternative to machine learning is crowd sourcing classifications of potential transient detections. *Galaxy Zoo Supernovae* (Smith et al. 2011) and *Snapshot Supernova* (Campbell et al. 2015) are two projects to have taken this approach, for the Palomar Transient Factory (PTF) (Rau et al. 2009; Law et al. 2009) and SkyMapper respectively. Both projects were facilitated through the *Zooniverse* Citizen Science platform [XXXXX Cite XXXXX]. Both projects asked volunteers to assess the target, reference and difference images for each detection and answer a series of questions that lead to a classification of real or bogus. The Galaxy Zoo Supernova team showed that it was possible to take the citizen science classifications and train a machine that outperformed volunteers in terms of a trade-off between purity and completeness. [XXXXXX I don't have a citation for this only a blog post. It seems that Bloom et al. 2012 didn't use GZ Supernovae classifications from what I can gather from a footnote in the paper... XXXXXX].

For classification tasks humans and machines have complimentary strengths - humans are good at making abstract connections given a small number of examples that they can apply in general while machines can consume large quantities of data and make more systematic judgements based on complex relationships between the features provided. Humans can quickly learn that magnitude has little influence or a real or bogus classification, while a machine must learn this from the data. But a machine can learn the typical relationship between the Point Spread Function (PSF) and mag-

⋆ E-mail: darryl@zooniverse.org

nitude to make systematic (though probably biased) classifications whereas humans are likely to use this information differently and in a more XXXX haphazard XXXX way. It remains to be seen whether classifications from a machine can be successfully combined with classifications from citizen scientists to produce more pure and complete samples from astronomical data. Of course we expect that combining machine classifications with those of experts will no doubt improve performance, the assumption here is that a small number of experts can not review all the data promoted by a machine classifier - the problem expected from LSST. It is unclear if the classifications of citizen scientists, which tend to be noisier can be added to improve the overall performance or if they are limited to providing a small but pure sample of the target class.

In this paper we report some initial findings from the *Supernova Hunters* project, a new citizen science project similar in spirit to those mentioned above but applied to the PanSTARRS Survey for Transients (PSST). In Section 2 we describe the PanSTARRS-1 telescope, PSST survey and the Supernova Hunters project and the citizen science platform. Section 3 shows the relative performance of humans and machines on data uploaded to Supernova Hunters during the first two months of the project. We also describe and measure the performance of a simple method for combining the classifications of citizen scientists and the current PSST machine classifier. In Section 4 we conclude and discuss potential avenues for future improvements.

2 METHOD

2.1 Pan-STARRS1

Pan-STARRS1 comprises a 1.8m primary mirror (Kaiser et al. 2010) and 60 4800 pixel detectors, constructed from $10\mu m$ pixels subtending 0.258 arcsec (Magnier et al. 2013) and a field-of-view of 3.3 deg. The filter set consists of g_{P1}, r_{P1}, i_{P1}, z_{P1} (similar to SDSS *griz* (York et al. 2000)), y_{P1} extending redward of z_{P1} and the "wide" w_{P1}-band filter extending over g_{P1} to i_{P1} (Tonry et al. 2012). Between 2010 and 2014 Pan-STARRS1 was operated by the PS1 Science Consortium (PS1SC) performing 2 major surveys. The Medium Deep Survey (MDS) was allocated 25% of observing time for high cadence observations of the 10 Medium Deep fields and the 3π survey allocated 56% observing time to observe the entire sky north of -30 degrees declination with 4 exposures per year in each of g_{P1}, r_{P1}, i_{P1}, z_{P1} and y_{P1} for each pointing.

The 3π survey was completed in mid-2014 and since then the telescope has been carrying out a NASA funded wide-field survey for near earth objects through the NEO Observation Program operated by the Pan-STARRS Near Earth Object Science Consortium (PSNSC). The NASA PSNSC survey is similar to the 3π survey optimised for NEO discoveries. Observations are in w_{P1} in dark time and combinations of $i_{P!}$, z_{P1} and y_{P1} during bright time. The Pan-STARRS Survey for Transients (PSST) (Huber et al. 2015) searching for the data for static transients and releasing these publicly within 12 to 24 hours.

Typically a single field is imaged 4 times in a night with exposures separated by 10-20mins called Transient Time Interval (TTI) exposures to allow for the discovery of moving objects. The quads of exposures are not dithered or stacked, meaning that cross-talk ghosts, readout artefacts and problems of fill-factor are inherent in the data (see Denneau et al. (2013) for some examples). Individual exposures are differenced with the 3π all-sky reference stack and sources in the resulting difference images are catalogued.

A series of pre-ingest cuts are performed before the catalogues are ingested into a MySQL database at Queen's University Belfast (QUB). These cuts are based on the detection of saturated, masked or suspected defective pixels within the PSF area in addition to flag checks for ghost detections and rejecting detections within ± 5 degrees galactic latitiude. Detections passing these cuts are grouped into transient candidates if they are spatially coincident within 0.5 arcsec and the rms scatter is < 0.25 arcsec. Post-ingest cuts are applied on detection quality, convolution checks and a check for proximity to brights objects. Additional cross-talk rules have been identified and implemented at QUB to reject ghosts not flagged at the pre-ingest stage. Remaining detections are cross-matched with the Minor Planets Centre Ephemeris database [XXXXX citation XXXXX] to identify any asteroids not removed by the rms cut. Remaining transient candidates are passed to our machine classifier described in the next section.

2.2 Convolutional Neural Network

In [XXXX Wright 2016 in prep. XXXXX] we show that the approach to training a machine classsifier described in Wright et al. (2015) was found to be inadequate for PSST, likely due to the larger variety of artefact types (a consquence of differencing individual exposures) and the expense of obtaining a large sample of labelled training data. Instead we turned to Convolutional Neural Networks (CNNs) that maintain the advantages of operating solely on the pixel data but at a higher computational cost in deployment.

The training set for this classifier was drawn from 3π survey data between 1st June 2013 and 20th June 2014. The sample of real detections are taken from spectroscopically confirmed real transients or that have been labelled by a human as high probability real detections and bogus detections a random subsample of detections discarded by post-ingest cuts or human eyeballing. The training set consists of 6916 examples with an additional 2303 detections held out for testing with both data sets containing twice as many bogus detections to real. Each example was manually inspected in order to limit label contamination; not all detections associated with a spectroscopically confirmed transient are necessarily real for example.

Given the small data set, to avoid overfitting we limit the the CNN to single convolution layer with 400 kernels and a pooling layer followed by a binary softmax classifier. We also perform unsupervised pre-training with sparse filtering Ngiam et al. (2011) using unlabelled images from the STL-10 (Coates et al. 2011) data set. The classifier is applied to nightly PSST data producing a score for each TTI exposure for every candidate passing the cuts in the previous section. The score or *hypothesis* is a function $h(x)$ of the input feature representation, x (the output of the convolution and pooling layers). For each candidate we simply combine the TTI exposure hypotheses by taking the median, leav-

ing a single real-bogus factor for each transient candidate which we take as the machine equivalent of $P(real|x)$ below. Candidates scoring ≤ 0.436 are automatically rejected with the remaining objects promoted for human screening. The decision boundary on $h(x)$ was chosen such that for a 5% false positive rate (FPR) we expect a missed detection rate (MDR) of $\sim 5.2\%$ based on the test set.

2.3 Citizen Science Platform

CL

Supernova Hunters was launched on the 12th July 2016 (MJD 57581). As of the 3rd September 2016 the project has accumulated 426481 classifications from 3158 citizen scientists with a few tens of "super users" submitting thousands of classifications. Citizen scientists are presented with the interface shown in Figure. 1 and have classified 52291 individual subjects corresponding to TTI observations (see Section 2.1) of 18838 PS1 objects. As guidance we provide a "Field Guide" that provides a desciption and examples of the different artefact types we expect. Every Tuesday $\sim$ 5200 new subjects are uploaded to the project consisting of the previous week's detections that pass our machine cuts. We require at least 7 citizen scientist classifications before a subject is considered classified and subsequently retired from the project. Since launch the project averages $\sim$ 23000 classifications in the first 24 hours after the data is released and $\sim$ 8300 classifications in the following 24 hours by which time all subjects are normally retired. High confidence (typically $P(real|x) > 0.8$) supernova candidates are screened by experts to remove a small number of false positives before the targets are submitted to the Transient Name Server (TNS). To date citizen scientists have discovered two hundred supernova candidates submitted to the TNS and two confirmed Supernovae including SN 2016els; a superluminous supernova Type I [XXXXX cite Gal-Yam?? XXXXX]. The classification spectra were obtained by PESSTO [XXXXX cite Smartt XXXXX].

3 PERFORMANCE

We present the results of our machine classifier (Section 2.2) in Figure 2 on PS1 data between MJD 57570 and MJD 57586. A major contaminant is the presence of asteroids; they appear in the difference image as supernovae but are identified here via cross-matching with the Minor Planet Centre ephemaris database (XXXXXX).

These data were additionally reviewed by at least one expert members of the team (normally DW or KS) to identify genuine supernovae. Candidates were divided into 'real' and 'bogus' categories based on these expert classifications.

Candidates with high probability as assigned by the machine are more likely to be real. However, although the machine successfully rejects the majority of bogus candidates, the sample produced by the simple cut on hypothesis is far from pure; 1403 real candidates from 3384 in the sample. Higher cutoffs run the risk of rejecting an increasing number of real candidates; requiring a 1% false positive rate will result in a missed detection rate of 60.3%.

An obvious solution to improve the performance of the neural network is to increase the size of the training set. This is expected to both directly improve performance but also allow for deeper, more complex networks to be trained. However, it is obvious already that obtaining large, clean training sets is expensive, requiring the review of many candidates by experts. In order to reduce the burden on the science team, candidates which exceed this threshold were also classified by volunteers via the *Supernova Hunters* project. The results of this analysis are shown in fig 3.

Volunteer classifications were combined using the simplest possible metric; the fraction of volunteers who identified a potential transient is assumed to be an estimate of the probability of that candidate being real. Despite this simple procedure, the results show that volunteers could effectively distinguish between real and bogus classifications. However, the structure of the resulting distribution is strikingly different. Whereas for machine classification, a threshold could be chosen to give a complete but not pure sample, with volunteer classification it is easier to construct a pure sample of candidates which are highly likely to be supernovae, but this sample is far from complete. There are candidates judged 'real' by experts even at low probabilities.

Much work has been done in other projects to improve on this naive combination of volunteer votes. (XXXX Examples XXXX). However, such solutions have not yet shown to be generalisable between projects, and so require substantial effort in data analysis. They may also depend on large numbers of volunteers classifying each subject. Instead of pursuing this route, therefore, we combine human and machine classifications hoping to benefit from the different capabilities of both sets of classifiers.

3.1 Combining human and machines

Figure 4 shows the combination of human and machine classifications. It is immediately apparent from the figure that no single threshold on either machine or human classification can outperform the combination of the two. This is an important result; it is the first time that the benefits of combining classification from both machines and volunteers has been clearly demonstrated using data from a real astronomical system.

How should the two independent classifications be combined? We simply apply a decision boundary of the form $\tau = \sqrt{P(real|x)^2 + h(x)^2}$ on the 2D surface, where $0 \leq \tau \leq 1$. For a constant value of τ a candidate is classified as bogus if $\sqrt{P(real|x)^2 + h(x)^2} <= \tau$ and classified real otherwise. This is equivalent to projecting the data onto $P(real|x) = h(x)$ producing a new scalar score for each detection (see Figure. 5). Figure. 5 shows the distribution of the resulting combined scores. This distribution maintains the purity at high combined scores that we see from human classifications, but also allows a better trade-off between purity and completeness. As an independent test we apply this same method to data between MJD 57587 and MJD 57627 in Figure. 6. Between MJD 57609 and MJD 57615 we relaxed our cut on $h(x)$ from 0.436 to 0.3 uploading any objects passing this cut to Supernova Hunters. This allowed the recovery of SN 2016fev at Type Ia supernova that would have been automatically rejected with $h(x)$=0.39, but which recieved a $P(real|x)$ of 1.0 from Supernova Hunters. The performance of the combination method on this data set is reported in tables 1 and 2 and the Reciever Operator Charateristic

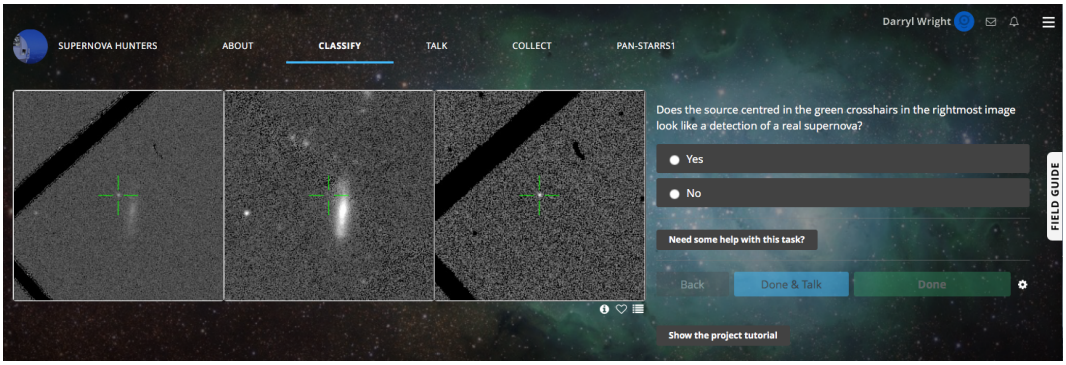

Figure 1. Screenshot showing the classification interface presented to citizen scientists. The left most image is the new *target* image taken recently. In the centre is the equivalent 3π *reference* image and on the right is the *difference* image. Volunteers are asked to decide whether or not they think the detection in the green crosshairs in the difference image is a detection of a real transient.

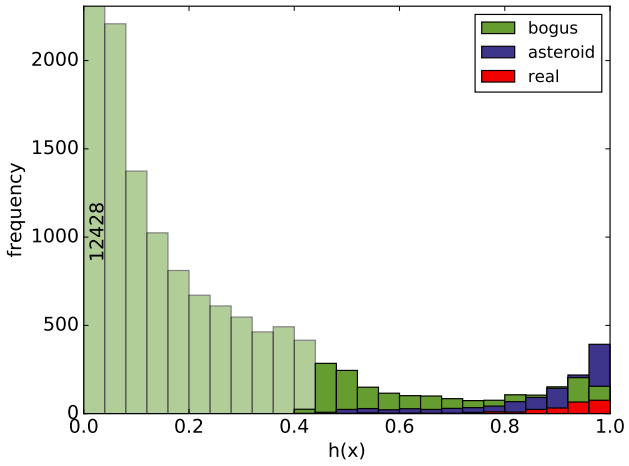

Figure 2. The distribution of hypotheses, $h(x)$ from the current 3π machine classifier for detected objects between MJD 57570 and MJD 57586. The light green shows the distribution of objects with $P(real) \leq 0.436$ which are automatically rejected. The remaining objects promoted for human screening even at high values of $h(x)$ contains many false positives.

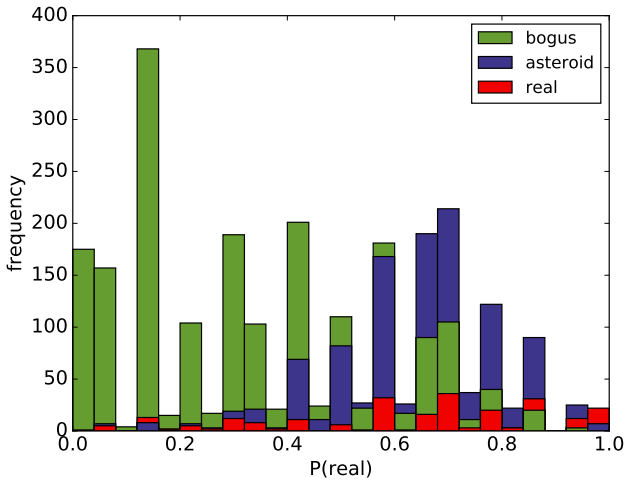

Figure 3. The distribution of $P(real)$ from Supernova Hunters for objects detected between MJD 57570 and MJD 57586. Compared with the machine $P(real)$ in 2 the objects at the extremes are pure. There are very few real detections with $P(real) < 0.04$ and few bogus detections above 0.92.

False Positive Rate	Human	Machine	Combination
1%	73.9%	90.1%	58.7%
5%	56.3%	69.7%	35.8%
10%	45.6%	46.7%	23.8%

Table 1. Missed detection rate recorded for a choice of false positive rates, based on expert classifications.

Missed Detection Rate	Human	Machine	Combination
1%	92.5%	85.9%	69.3%
5%	75.1%	52.8%	41.8%
10%	53.8.8%	39.1%	26.5%

Table 2. False positive rate recorded for a choice of missed detection rates, based on expert classifications.

(ROC) Curve and Purity-Completeness (Precision-Recall) Curves are shown in Fig. 7.

For any choice of false positive rate, the combination of classifications produced a lower missed detection rate. Equally, for any required purity or completeness the combination provides a better trade-off.

We have chosen to implement one of the simplest methods for combining human and machine classifications to demonstrate how they complement one another. But it is easy to think of more complex combination methods, for example we trained a linear SVM on the data presented in Fig. 4 and found, unsurprisingly, that the performance measured on the data in Fig. 6 was typically within 1% of the values reported in tables 1 and 2. Although the gains in this example are negligible, if we wish to incorporate additional information they become important. As an example we also have a machine score from a classifier trained on catalogue information (similar in concept to the features of Bloom et al. (2012), Brink et al. (2013) and Goldstein et al. (2015)) for each of the Supernova Hunters detections. It is

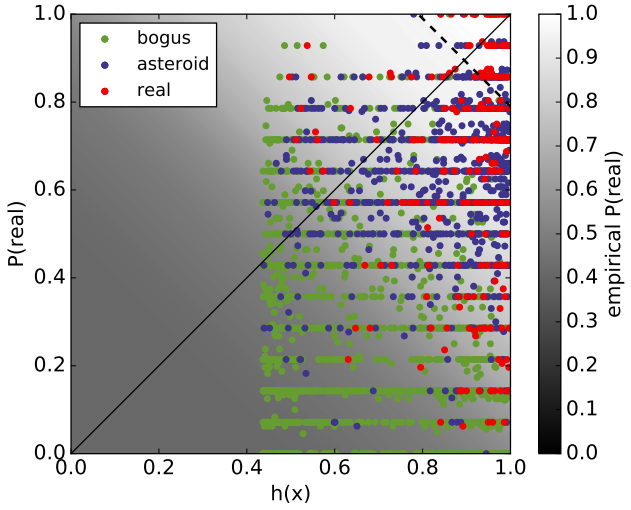

Figure 4. The $P(real)$ from Supernova Hunters against the machine $P(real)$ for detected objects between MJD 57570 and MJD 57586. $P(real)$ and $h(x)$ are combined by projecting the data onto the solid black line in the euclidean sense. For a given value of τ the background colour map shows the probability that an example chosen at random with combined score abouve tau will be real; an empirical measure of P(real) for our combination of method. The dashed black line shows the 90% probability contour.

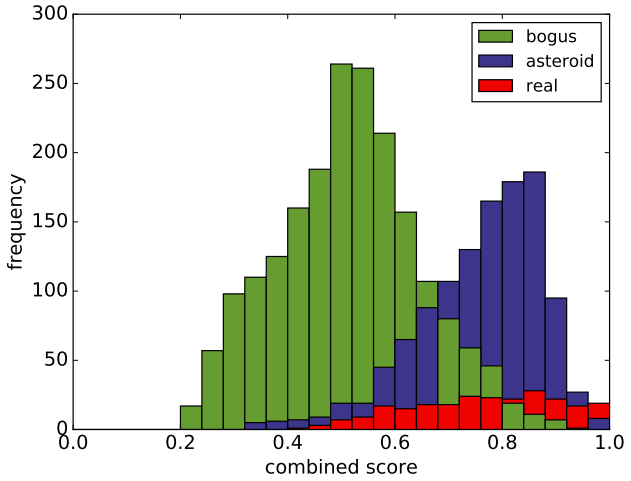

Figure 5. Histogram showing the distribution of data resulting from the combination of human and machine classifications.

unlikely that the combination method presented here will work for this higher dimensional representation, but we can expect that an SVM may gain from the added information.

4 CONCLUSIONS

CJL

In this paper we introduced a new citizen science project, Supernova Hunters built entirely with the XXXXX off-the-shelf XXXXX Zooniverse Project Builder requiring no custom features or additional development. The project

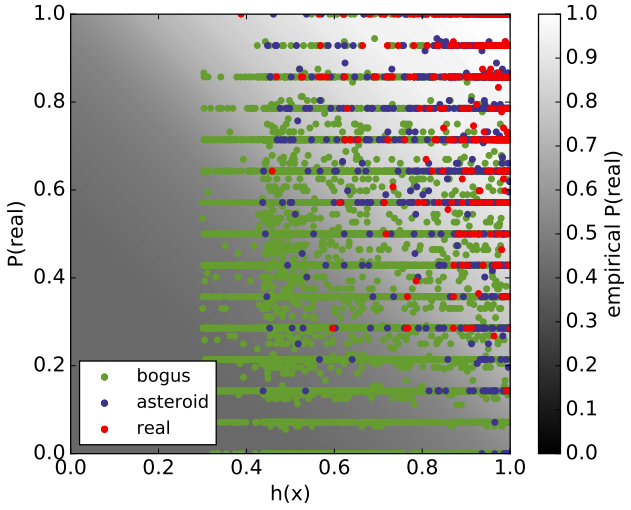

Figure 6. The same as 4 but on a new sample of 10908 objects detected between MJD 57587 and MJD 57627. Between we relaxed our cut on $h(x)$ to 0.3 which allowed us to recover a supernova with $h(x)$=0.39, but achieved a $P(real)$=1.0 from Supernova Hunters. The background colour map is the same empirical $P(real)$ as Figure. 4 but underestimates the probability at each value of τ for this data set, perhaps suggesting an improvement in the classification ability of volunteers over time.

aims to classify detections of potential supernova candidates from the PanSTARRS Survey for Transients as either real transients or bogus detections of instrumentation or image processing artefacts. To be uploaded to the Supernova Hunters project a detection must first pass a series of cuts based on catalogue information and secondly be promoted by our machine learning algorithm, namely a Convolutional Neural Network. With this approach we expect that only about 5% of the false positives passing the catalogue cuts make it into the project, greatly reducing the number of objects we ask volunteers to screen. Citizen scientists excel at mining this data for a very pure sample of high $P(real|x)$ supernova candidates, typically those of higher signal-to-noise and offset from a galaxy. But compared with expert labels many less obvious candidates are missed. Rather than consider methods of weighting classifications of individual citizen scientists based on past performance on gold standard data, we instead applied a simple combination of the scores provided by humans and machines. We showed the new combined score achieved better performance than either individually for any choice of purity or completeness.

We expect that there are many ways to improve on the work presented here. Improving the ability of both humans and machines to discriminate between classes is likely to improve the combination. For example we could invest more effort into educating citizen scientists to identify more subtle artefact indicators, there are many bogus detections with $P(real|x) > 0.5$, but relatively few real detections below 0.5 in Figure 3. The current machine classifier was trained at the beginning of the PSST survey and the algorithm was specifically chosen to learn from the limited amount of training data available. Given the large volume of data accumulated since we could train more sophisticated algo-

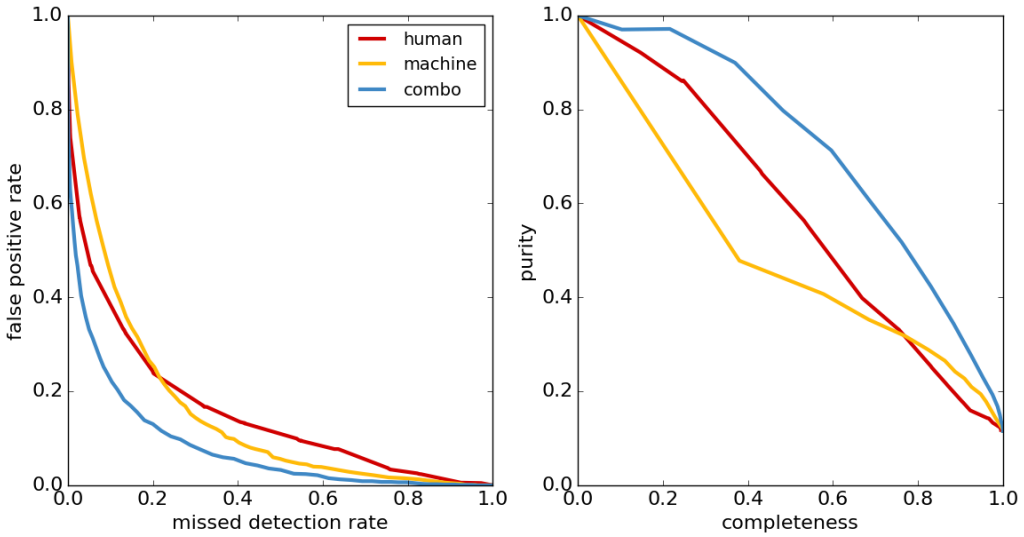

Figure 7. left: ROC curve showing performance measured on the test data in 6 for human (red), machine (yellow) and the combination of human and machine classifications (blue). right: The equivalent Purity-Completeness curve. Both plots show that the combination always outperforms humans or the machine individually.

rithms that can learn more complex relationships between the features, though extracting robust labels for this additional data is a challenge that still needs to be addressed. Additionally, we could look at generalising existing classifier weighting schemes such as SWAP (Space Warps Analysis Package) (Marshall et al. 2016) to improve $P(real|x)$ or as another method for combining humans and machine classifications.

This offers hope for dealing with the large data volumes from all sky surveys such as LSST and ATLAS. We used machines to reject the vast majority of false positives and then combining the machine hypotheses with classifications from a few thousand citizen scientists for the remaining candidates. With Supernova Hunters we have not actively sought additional citizen scientists, beyond the $\sim$ 30000 volunteers on the Zooniverse beta testing e-mail list, who were asked to review the project before launch, volunteers must "discover" the project on the Zooniverse projects page to participate. Given that we could actively seek the participation of $\sim 10^6$ registered Zooniverse volunteers and assuming that 10% chose to participate, given the current classification rate we could achieve $\sim$ 750000 classifications per night, or 0.75 classifications for the $\sim 10^6$ transient alerts expected from LSST at the beginning of the survey (Ridgway et al. 2014). If the false positive rate is an order of magnitude (perhaps overly pessimistic given the expected $\sim$ 500-2200 false positives per field per visit (Ridgway et al. 2014)) more than the transient alert rate and assuming we can discard 90% of those with machine learning we can expect to achieve 0.375 citizen science classifications per promoted detection. If 10 classifications are required per detection before considering it classified, we are roughly two orders of magnitude short. With continued improvements to difference imaging (Zackay et al. 2016), automated real-bogus classification and more efficient use of citizen scientist classifications making up this deficit maybe achievable.

REFERENCES

Bailey S., Aragon C., Romano R., Thomas R. C., Weaver B. A., Wong D., 2007, ApJ, 665, 1246
Bloom J. S., et al., 2012, PASP, 124, 1175
Brink H., Richards J. W., Poznanski D., Bloom J. S., Rice J., Negahban S., Wainwright M., 2013, MNRAS, 435, 1047
Campbell H., et al., 2015, The Astronomer's Telegram, 7254
Coates A., Lee H., Ng A. Y., 2011, in AISTATS 2011.
Denneau L., et al., 2013, PASP, 125, 357
Donalek C., Mahabal A., Djorgovski S. G., Marney S., Drake A., Glikman E., Graham M. J., Williams R., 2008, in Bailer-Jones C. A. L., ed., American Institute of Physics Conference Series Vol. 1082, American Institute of Physics Conference Series. pp 252–256 (arXiv:0810.4945), doi:10.1063/1.3059057
Goldstein D. A., et al., 2015, AJ, 150, 82
Huber M., et al., 2015, The Astronomer's Telegram, 7153
Kaiser N., et al., 2010, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. , doi:10.1117/12.859188
Law N. M., et al., 2009, PASP, 121, 1395
Magnier E. A., et al., 2013, ApJS, 205, 20
Marshall P. J., et al., 2016, MNRAS, 455, 1171
Ngiam J., Chen Z., Bhaskar S. A., Koh P. W., Ng A. Y., 2011, in Shawe-Taylor J., Zemel R., Bartlett P., Pereira F., Weinberger K., eds, , Advances in Neural Information Processing Systems 24. Curran Associates, Inc., pp 1125–1133, http://papers.nips.cc/paper/4334-sparse-filtering.pdf
Rau A., et al., 2009, PASP, 121, 1334
Ridgway S. T., Matheson T., Mighell K. J., Olsen K. A., Howell S. B., 2014, ApJ, 796, 53
Romano R. A., Aragon C. R., Ding C., 2006, in Machine Learning and Applications, 2006. ICMLA'06. 5th International Conference on. pp 77–82

Smith A. M., et al., 2011, MNRAS, 412, 1309
Tonry J. L., et al., 2012, ApJ, 750, 99
Wright D. E., et al., 2015, MNRAS, 449, 451
York D. G., et al., 2000, AJ, 120, 1579
Zackay B., Ofek E. O., Gal-Yam A., 2016, preprint, (arXiv:1601.02655)
du Buisson L., Sivanandam N., Bassett B. A., Smith M., 2015, MNRAS, 454, 2026