Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **499,** 6009–6017 (2020)
Advance Access publication 2020 October 9

doi:10.1093/mnras/staa3096

Machine learning for transient recognition in difference imaging with minimum sampling effort

Y.-L. Mong,[1,2]⋆ K. Ackley,[1,2] D. K. Galloway,[1,2] T. Killestein,[3] J. Lyman ,[3] D. Steeghs,[3] V. Dhillon ,[4]⋆ P. T. O'Brien,[5] G. Ramsay,[6] S. Poshyachinda,[7] R. Kotak,[8] L. Nuttall,[9] E. Pallé,[10] D. Pollacco,[3] E. Thrane,[1] M. J. Dyer ,[4] K. Ulaczyk,[3] R. Cutter ,[3] J. McCormac,[3] P. Chote,[3] A. J. Levan,[3] T. Marsh,[3] E. Stanway ,[3] B. Gompertz ,[3] K. Wiersema,[3] A. Chrimes ,[3] A. Obradovic,[1] J. Mullaney,[4] E. Daw,[4] S. Littlefair,[4] J. Maund ,[4] L. Makrygianni,[4] U. Burhanudin,[4] R. L. C. Starling ,[5] R. A. J. Eyles-Ferris,[5] S. Tooke,[5] C. Duffy,[6] S. Aukkaravittayapun,[7] U. Sawangwit,[7] S. Awiphan,[7] D. Mkrtichian,[7] P. Irawati,[7] S. Mattila,[8] T. Heikkilä,[8] R. Breton ,[11] M. Kennedy ,[11] D. Mata Sánchez[11] and E. Rol[1,2]

[1] *School of Physics and Astronomy, Monash University, Clayton, VIC 3800, Australia*
[2] *OzGrav: The ARC Centre of Excellence for Gravitational Wave Discovery, Clayton, VIC 3800, Australia*
[3] *Department of Physics, University of Warwick, Coventry, West Midlands CV4 7AL, UK*
[4] *Department of Physics and Astronomy, Hicks Building, The University of Sheffield, Sheffield S3 7RH, UK*
[5] *School of Physics and Astronomy, University of Leicester, University Road, Leicester LE1 7RH, UK*
[6] *Armagh Observatory and Planetarium, College Hill, Armagh BT61 9DB, UK*
[7] *National Astronomical Research Institute of Thailand, 260 Moo 4, T. Donkaew, A. Maerim, Chiangmai 50180, Thailand*
[8] *Department of Physics and Astronomy, University of Turku, FI-20014 Turku, Finland*
[9] *Institute of Cosmology and Gravitation, University of Portsmouth, Dennis Sciama Building, Burnaby Road, Portsmouth PO1 3FX, UK*
[10] *Instituto de Astrofisica de Canarias, La Laguna, E-38205 Tenerife, Spain*
[11] *Department of Physics and Astronomy, The University of Manchester, Oxford Road, Manchester M13 9PL, UK*

Accepted 2020 October 5. Received 2020 September 11; in original form 2020 June 17

ABSTRACT

The amount of observational data produced by time-domain astronomy is exponentially increasing. Human inspection alone is not an effective way to identify genuine transients from the data. An automatic real-bogus classifier is needed and machine learning techniques are commonly used to achieve this goal. Building a training set with a sufficiently large number of verified transients is challenging, due to the requirement of human verification. We present an approach for creating a training set by using all detections in the science images to be the sample of real detections and all detections in the difference images, which are generated by the process of difference imaging to detect transients, to be the samples of bogus detections. This strategy effectively minimizes the labour involved in the data labelling for supervised machine learning methods. We demonstrate the utility of the training set by using it to train several classifiers utilizing as the feature representation the normalized pixel values in 21 × 21 pixel stamps centred at the detection position, observed with the Gravitational-wave Optical Transient Observer (GOTO) prototype. The real-bogus classifier trained with this strategy can provide up to 95 per cent prediction accuracy on the real detections at a false alarm rate of 1 per cent.

Key words: methods: data analysis – methods: statistical – techniques: image processing.

1 INTRODUCTION

Transient astronomy focuses on astrophysical objects that vary on timescales of hours to years, and can originate from events such as supernovae, accreting binaries, stellar flares, tidal disruption events, and gamma-ray bursts (GRBs). Identifying and characterizing transients is important for understanding astrophysics under extreme environments, accretion physics, and the underlying physics of stellar flares.

In 2015, transient science stepped into a new era with the first direct detection of a gravitational wave (GW) event, GW150914 (Abbott et al. 2016), caused by the merger of a pair of ≈30-$M_\odot$ black holes. Two years later, the first binary neutron star merger, GW170817, was detected (Abbott et al. 2017). GW detection alone can typically localize the event to only within a few hundred square degrees. To improve the localization down to order of an arcsecond, rapid-response electromagnetic follow-up observations are required (e.g. Coulter et al. 2017). The identification of electromagnetic counterparts to the GW events is key to understanding the environments of the post-merger remnants (Metzger 2017).

All-sky optical surveys can provide a more complete investigation of the optical transient sky. Time-domain astronomy has become a fast-growing area of astrophysics requiring comprehensive rapid-responsive strategies for following up the triggers of interesting events, such as GRBs and GW events.

⋆ E-mail: yik.mong@monash.edu

The recent advances of transient astronomy have been well established by many transient survey projects, such as the SDSS-II Supernova Survey (Frieman et al. 2008), the Catalina Real Time Transient Survey (CRTS; Drake et al. 2009), Pan-STARRS1 (PS1; Kaiser et al. 2010), the Zwicky Transient Facility (ZTF; Masci et al. 2018), the Asteroid Terrestrial-impact Last Alert System (ATLAS; Tonry et al. 2018), and the SkyMapper Transient Survey (Wolf et al. 2018), among others. In the future, of order 10^6 transients are expected to be discovered per night with the *Vera C. Rubin Observatory* (Ivezić et al. 2019).

The Gravitational-wave Optical Transient Observer[1] (GOTO) is a robotic ground-based optical telescope located at the Roque de los Muchachos Observatory on La Palma, Canary Islands (Steeghs et al., in preparation). It is dedicated to searching for the optical counterparts to GW events. The GOTO prototype currently consists of 4 × 40 cm unit telescopes (UTs) covering $\approx 18\,\mathrm{deg}^2$ per exposure. The angular resolution of GOTO is about $1.24\,\mathrm{arcsec\,pixel}^{-1}$. There are four Baader filters on each UT, a broad-band L filter (400–700 nm), and narrower B, G, and R filters. Under dark conditions, the detection limit in the L band is ≈ 20.5 mag in three stacked 60-s exposures. GOTO also performs an all-sky survey in order to discover other types of optical transients. The nightly sky coverage of GOTO is up to $\approx 2000\,\mathrm{deg}^2$.

To detect transients in an all-sky survey, difference imaging and ‘real-bogus’ classification are the key steps. Difference imaging is the process under which a recently observed ‘science’ image is subtracted from an earlier ‘reference’ image for identifying excess flux (see Section 2.2 for more details). However, as the difference images include both subtraction residuals and transient detections, real-bogus classification is required to separate them.

Due to a large number of detections (typically $\gtrsim 10^4$) per GOTO image, source vetting and identification cannot rely solely on manual inspection. Efficient ‘real-bogus’ classification on difference images has become one of the most important problems in transient astronomy, and several techniques have already been developed based on both supervised and unsupervised machine learning to address the problem.

There are two traditional ways to extract feature representations using supervised machine learning. Isophotal measurements of the detections (hereafter referred to as ‘level-0’ attributes) could be used as the model features (Bloom et al. 2012; Brink et al. 2013; Gieseke et al. 2017). Additionally, both linear and non-linear combinations of level-0 attributes could generate more useful, but complicated features (hereafter referred to as ‘level-1’ attributes). However, there are a vast number of ways to combine level-0 attributes, and trial and error tests have to be carried out in order to verify which level-1 attributes are useful. This ‘feature engineering’ step becomes the most challenging part of the method. On the other hand, Wright et al. (2015) and Gieseke et al. (2017), hereafter referred to as W15 and G1, respectively, use pixel intensities as the feature representatives, which do not require any feature engineering.

Previous studies have shown that the learning algorithm and size of the training data set are the key factors affecting the performance of the classifier. W15 used a sample size of 32 095 (80 per cent training data, and 20 per cent test data). Brink et al. (2013), on the other hand, trained their classifiers on 50 000 detections and tested the classifiers on a validation set with a size of 28 448. The random forest (RF) technique is a machine learning algorithm with the architecture of multiple decision trees. It performed best in terms of the FOM for both W15 and G17 studies, i.e. using either isophotal measurements or normalized pixel values as the classification features. The FOM is defined as the minimum missed detection rate (MDR) with an acceptance of 1 per cent false positive rate (FPR). A convolutional neural network (CNN) is another machine learning algorithm that is now widely used for image recognition in many different fields. Unlike RF, CNN only adopts pixel values to be the learning features. Some authors (e.g. Cabrera-Vives et al. 2016, 2017; Gieseke et al. 2017) have claimed that CNN shows the best performance at picking out real candidates in difference images.

The most challenging part of applying supervised machine learning is in building up a sufficiently sized training data set in an automated way. Relying on human classification alone to create the training set is prohibitively expensive. Real transients in the data set could be defined as known transients identified by archival catalog searches or with prompt spectroscopy. Wright et al. (2015) built up a data set of ≈ 8000 real transients based on 3 yr of Pan-STARRS1 observations, while Brink et al. (2013) used PTF observations taken in 2010 to build their training data set, where they identified 14 781 real transients on difference images based on spectroscopy and other public domain data to create their real sample.

In this paper, we describe how we build a real-bogus classifier with minimum sampling effort. We begin with the motivation of this work followed by a brief description of the image processing in Section 2. We describe the construction of our data sets and the feature extraction in Sections 3 and 4. The models we use are described in Section 5. In Section 6, we compare the performance of our ‘quick-build’ classifier (QB-classifier) with the one trained on an injection set (IT-classifier). Finally, we summarize our work in Section 7.

2 MOTIVATION AND IMAGE PROCESSING

2.1 Motivation

The most straightforward approach to building a sample of real transients in the training set is to manually separate these from the few thousand bogus detections in each difference image (see Section 2.2 for more details on difference imaging). Using information from other transient surveys with spectroscopic classification, we can ensure that our sample of real transients is pure. However, there are two main problems with this approach. First, each of the samples in the data set has to be classified manually, which is a labour-intensive exercise. Secondly, it takes a long time to build a large data set, and the exercise is not easily scalable to even larger data sets.

To solve these problems, we have to understand how real detections appear on difference images. Unlike real detections, bogus artifacts typically do not appear as point sources in the difference images. Consequently, one can reasonably assume that genuine transients in the difference images should have similar properties to the point sources in the science images, since both detections can be described by a PSF superimposed on top of background noise. We can therefore create our training data set by collecting the training sample from the detections on the science images rather than from the difference images. This method of assembling a training data set does not require any human inspection allowing us to easily build up a very large sample.

There are several potential contaminants in the resulting sample: extended objects, such as galaxies, and artefacts, including cosmic rays, and hot pixels. The contaminant fraction can be reduced by filtering the outliers from the normalized full width at half-maximum (FWHM) distribution, and by using the `SExtractor` parameters

[1] https://goto-observatory.org/.

Table 1. Number of detections in different data sets.

Data sets	Bogus	Real	Total
Quick-build training set	400 000	400 000	800 000
Injection data set	141 782	141 782	283 564
MP test set	42 929	33 511	76 440

CLASS_STAR and *ISO* AREA_IMAGE to exclude the galaxies and hot pixels from the real sample (see Section 3.1 for more details on SExtractor).

In parallel to the approach we used to build our real sample, we build our bogus sample by collecting all detections on the difference images. Since we label all detections on the difference images as bogus, there may be some genuine transients included in the bogus sample. The contamination fraction in the bogus sample is estimated to be less than 1 per cent by assuming no more than 20 transients on each field.

With a large training data set, the machine learning model is less likely to be overfitted. Therefore, the decision boundary should be smooth enough to reject the outliers, which are the contaminants in our training set. As a result, we can maintain this negligible contamination in both real and bogus samples.

The key aim of this work is to demonstrate that our method of creating the training set is not only effective, but also easily applicable to different machine learning algorithms to solve the real-bogus classification problem. We have therefore implemented different algorithms into the classifier to verify the feasibility of our approach.

2.2 Image processing

Raw images taken with GOTO are reduced automatically with our standard pipeline before performing further analysis (Ackley et al., in preparation). The standard pipeline applies bias correction, dark-frame subtraction, and flat-field correction, followed by co-adding 3 × 60 s individual exposures to form a median science image. Throughout this study, we performed all analyses using median images as these have a higher signal-to-noise ratio than individual exposures.

The template image, also referred to as the reference image, is a previous image of the same field that is subtracted from all successive science frames. Since GOTO operates by tiling the sky on a fixed grid (Dyer et al. 2018), we are able to update the set of templates regularly.

Image alignment and difference imaging are part of the standard pipeline procedures following calibration. We use a modified version of the PYTHON package alipy to align the template image to the science image by cross-matching positions of selected field stars using high-order affine transformations independent of the WCS information. Once the alignment has been performed, we use hotpants[2] (Becker 2015) to perform image subtraction.

3 DATA SETS

We use three data sets in this work: the quick-build training set, the injection data set for both testing and training, and the minor planet (MP) test set (Table 1).

The quick-build training set is used to train our real-bogus classifier. We apply our quick-build strategy which can effectively assemble real detections in our training set. In practice, we are primarily concerned about the performance of the classifiers applied on the difference images. Since all the real samples in this training set are collected from the science images, this data set will not be used for any testing purpose. Therefore, as we need a reliable test set for testing the performance of our classifiers, we are motivated to build the injection data set and MP test set.

The injection data set is generated by collating all of the detections from the difference images after performing difference imaging on the injected science images. We apply this data set in two ways. The first is to use the injection set to test the classifiers trained on the quick-build training set. On the other hand, since the morphology of the injections are a good representative of how genuine transients may appear in practice, we also use the injections recovered on the difference images to train our classifier. This, in effect, mimics the training process using genuine transients in the standard way and will indirectly provide a figure of merit (FoM) comparison with the classifiers trained using the quick-build method.

We consider the MP test set to be the most reliable test set, over the injection test set, as we use verifiable MPs as our real sources. The classifiers trained on the quick-build and the injection data sets will be tested on this MP test set for performance evaluation and to provide evidence of the efficacy of our method.

3.1 'Quick-build' training set

To ensure that the quality of the images used to build our data set is sufficiently high enough, we select images based on several criteria. We randomly select 45 science images between 2019 April and May from different fields taken with different UTs for building our real sample. We avoid choosing images where the number of detections are <15 000 within the FoV ($= 2.1 \times 2.8\,\mathrm{deg}^2$) of a single UT to ensure a large enough representation of samples.

SExtractor is commonly used to identify detections, which have a higher pixel counts as compared to the background level, in an image (Bertin & Arnouts 1996). The default sensitivity parameter of the SExtractor, DETECT_THRESH, is set to 2.0σ for science images and 2.5σ for difference images on GOTO standard pipeline.

In order to avoid bad detections in our real sample, such as those on the edge of the frame, saturated or spurious pixels, etc., we filter out the detections with non-zero FLAGS.[3] This step will remove saturated bright objects (FLAG=4), and any objects that are too close to bright objects (FLAG=2). For those objects next to bright objects that are well deblended (FLAG=0), they are also included in our training set since they should still resemble a PSF on top of the background. We identify that flagged detections contribute ≈10 per cent of the entire real sample. We further reduce the contaminants by filtering the detections falling outside the range between 0.3 and 99.7 per cent percentiles of the normalized FWHM distribution over each image, as well as detections brighter than $m = 12$ were also removed in order to reduce the contamination due to bright objects with diffraction spikes. Finally, we build our real sample of 455 673 objects purely using the detections extracted from the science images.

Similarly, we use 680 775 detections extracted from 49 difference images to build our bogus sample. There is a small fraction of true negative contaminants in the bogus sample due to the presence of

[2] https://github.com/acbecker/hotpants.

[3] https://sextractor.readthedocs.io/en/latest/Flagging.html.

real transients in the difference image. In most cases, $\approx 10^3$–10^4 detections are recovered by SExtractor in a single difference image. Among them, there are typically fewer than 20 real transients per image, i.e. typically < 1 per cent. For those frames aligning on the galactic plane, there could be a higher number of recovered variable sources. However, the bogus artefacts that arise due to image subtractions residuals or template misalignments greatly outnumber the number of variable sources or true transients, and generally scale with the density of sources in the field. Therefore, the contamination fraction still remains less than 1 per cent. Building our bogus sample using all of the detections on the difference image (less than 1 per cent contamination) without human inspection is acceptable if the sample size is large enough. Combining with the real sample, our entire data set contains 1 136 448 detections.

To ensure our training set is balanced, we randomly select 400 000 detections from each of the real and bogus samples, to form our training set for a total size of 800 000 detections.

3.2 Injection data set

We create another data set by using images with simulated sources injected into them. We use this data set both for testing the performance of the classifiers trained on the quick-build training set (see Section 3.1), and to train another classifier for comparison purposes.

We use the field of SN2019pjv located at $\alpha = 17{:}14{:}34.8$, $\delta = +28{:}07{:}26.1$ (J2000), which has been revisited by GOTO 91 times on different nights between 2019 September and 2020 February, as our injection field. Since UT3 and UT4 were relatively stable, in terms of the FWHM compared to other UTs, we select images with QUALITY_FLAG = 0, for which the quality assessment of the images is calculated (see Ackley et al., in preparation), to perform injections, resulting in a total of 143 injected images.

We perform the injection process using iraf (Tody 1986, 1993). We uniformly inject point sources over each image, with apparent magnitudes in the range $m = 15$–21. The total number of injections which are recovered by SExtractor after difference imaging is 70 891, giving a 63 per cent recovery rate.

We define all 70 891 injections on the difference images to be the real sample of our training set. Furthermore, we double our real sample by reflecting all injection stamps along the diagonal image axis in order to create a larger data set. To build a balanced data set, we sample 141 782 bogus detections randomly for our bogus sample. In sum, our entire injection training set contains 283 564 detections.

3.3 MP test set

As a representative example of on-sky performance for genuine transient sources, we assemble a test set using archival MP detections from the past year of GOTO operations. This data set has the benefit over an injection set for accurately sampling across a wide range of field densities, image PSFs, and sky conditions. MP detections have similar properties to those of genuine transient objects – they are detected in the science image, but absent in the template image, due to the large sky motion of the object, which leaves a ‘clean’ subtraction residual and is similar to what we expect from genuine transient sources.

To build this test set, we randomly select 12 000 images from the GOTO data base. For each image, we obtain the positions of all known MPs within the field of view using the SkyBoT cone search (Berthier et al. 2006). These positions are then cross-matched with the difference photometry table of each image, to identify the detected MPs in each image. We adopt a cross-match radius of 1 arcsec, to minimize contamination from spurious associations. To generate a matching bogus sample for the test set, we randomly sample from the difference image detections, choosing one for every MP detected per image. This approach provides an unbiased sample of the typical bogus content of each image and, due to the significant imbalance between real and bogus detections, provides a largely clean bogus sample.

The largest source of contamination within this sample is variable stars. Inevitably, when selecting a random sample of sources in the image a small fraction of these will be variable, and could show a clear residual in the difference image, depending on the amplitude of variability. Those with clear residuals will have incorrect (bogus) labels and be marked as misidentifications in the training set due to the classifier scoring them as real. These contaminants would negatively skew any performance metrics calculated. Determining algorithmically which labels to assign these detections is difficult, and is likely to inject bias. We opt to remove all variable stars from the training set. After generating the test set, we cross-match the coordinates of the random bogus sample against the ATLAS Variable Star Catalogue (Heinze et al. 2018), with a generous cross-match radius of 5 arcsec. This aims to maximize completeness in removal of variable stars, at the cost of some non-variable objects being removed. Typically around 4 per cent of the test set is removed with this cut.

As a final cut, we remove cosmic rays from this test set. These features cannot always be distinguished in the difference image alone because when hotpants convolves the science image with the PSF kernel, these detections become PSF-like. We reject detections that only have one detection in the individual images that form the median. We opt for this approach to avoid removing sources that are undetected in the individual images due to poor signal-to-noise ratio, yet appear in the median stack.

Applying all of the steps above results in a test set of $\approx 76\,440$ examples, with the ratio 1: 1.6 MPs to random bogus detections. Our methodology for automated test set production is detailed more thoroughly in Killestein et al. (in preparation).

4 FEATURE EXTRACTION AND PRE-PROCESSING

To extract the pixel intensity features, we crop a 21×21 pixel (26×26 arcsec^2) stamp centred at the image coordinate (X_IMAGE, Y_IMAGE) of the detection as measured by SExtractor for each sample in the training set (see Fig. 1 for some examples). The real detections are all located at the centre of the stamp with a typical aperture size of ≈ 5 arcsec surrounded by shot noise. On the other hand, the segmentation of the subtraction residual might occur such that SExtractor would identify multiple bogus detections for a single astronomical object. The *red* framed bogus stamp in Fig. 1 is an example showing that a single object is segmented into three detections after difference imaging. It typically results in an offset between the segment of each subtraction residual and the actual position of the source.

Due to the appearance of masked pixels and missing values (off-edge pixels) within the pixel stamp in some cases, data cleaning was necessary before performing further analysis. During the subtraction of bright sources, masked pixels can be generated in the difference image (e.g. see the bottom left-hand thumbnail in Fig. 1). We clean the data by replacing all masked and off-edge pixels by the median value of the stamp, which is approximately the background level.

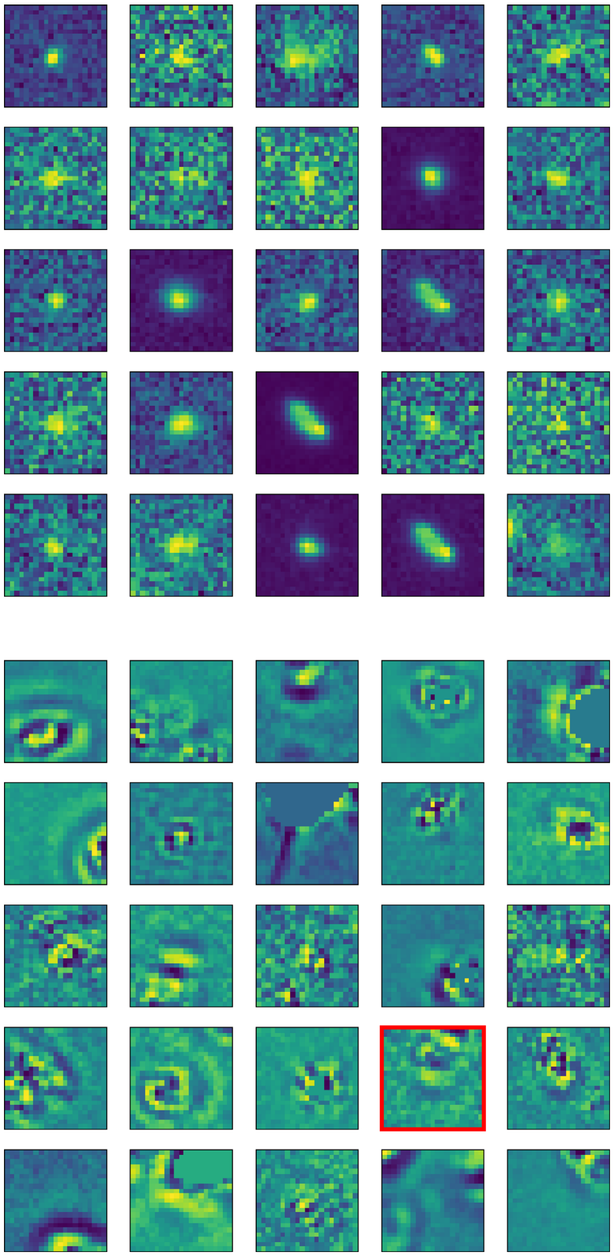

Figure 1. Examples of the 21 × 21 pixel thumbnails in the training set. Top five panels: examples of real detections. Bottom five panels: examples of bogus detections. The red framed bogus stamp shows the segmentation of detection in the image subtraction process.

Since each detection has its own signal-to-noise level relative to the background noise, we normalize the pixel intensities with

$$f(p_i) = \frac{p_i - \tilde{p}}{|p_i - \tilde{p}|} \log_{10}\left(1 + \frac{|p_i - \tilde{p}|}{\sigma}\right), \qquad (1)$$

where p_i is the ith pixel value. $\tilde{p}$ and σ are the median and the standard deviation of the pixel intensities in the stamp. This scaling algorithm is adopted from W15 and EYE[4] (Bertin 2001), with the modification that p_i is replaced by $p_i - \tilde{p}$. In previous studies (e.g. Bloom et al. 2012; Brink et al. 2013; Wright et al. 2015; Gieseke et al. 2017), the real sample was collected from the difference image, implying that the background level should always be around zero. In our case, since we use unsubtracted science image detection to comprise the real sample, the background level is always non-zero. Therefore, we reset the noise level at the median pixel value of the stamp.

Table 2. Model parameters of the ANN and the RF we adopt.

Model parameters	Values
Artificial neural network	
Size of first Dense layer	100
Activation (hidden layer)	ReLu
Regularization	$\lambda = 0.03$
Optimizer	RMSprop
Random forest regressor	
n_estimators	1000
max_features	25
min_samples_leaf	1

5 CLASSIFICATION ALGORITHMS

We build our classifiers using two different supervised machine learning algorithms: the RF (Breiman 2001) and the artificial neural network (ANN; McCulloch & Pitts 1943). These algorithms are selected due to their reasonable performance shown in the literature (Wright et al. 2015). We use the PYTHON packages sklearn (Pedregosa et al. 2011), keras, and tensorflow (Abadi et al. 2015) to build the RF and the neural network models, respectively.

We tune the hyperparameters of each model to optimize performance, and list the optimal hyperparameters in Table 2. We build our single-layer ANN model with 100 neurons. Activation functions ReLu and softmax are used in the hidden layer and the output layer respectively. The optimizer we use in ANN is RMSprop. For our RF classifier, we build it with n_estimators = 1000, max_features = 25, and min_samples_leaf = 1.

6 RESULTS AND PERFORMANCE

In this section, we show the general performance of the classifiers trained on the different data sets.

In order to mimic a more realistic case of applying our classifier to difference images directly, we verify the efficacy of different learning algorithms by testing on the injection data set (see Section 6.1). We also compare the performance between the classifiers trained on the quick-build training set and the injection data set. In Section 6.2, we compare the performance of the classifiers trained on different data sets by testing them on our MP test set.

6.1 Performance of the injection test

The injection data set consists of 283 564 samples with a 1:1 balance ratio between the numbers of real and bogus detections. We label all injections as real detections and leave the rest as bogus. Therefore, since there are some real transients existing on the difference images which are not injections but are labelled as bogus, the false-positive rate calculated from the injection test could be overestimated. With known magnitudes of all injections, we can study how the recovery rate would be affected by the brightness of the detection.

We compare the performance of ANN and RF models by plotting the receiver operator characteristic (ROC) curves (see Fig. 2). We conclude that the RF classifier performs better in terms of both area under the curve (AUC) and FoM.

[4] http://www.astromatic.net/software/eye.

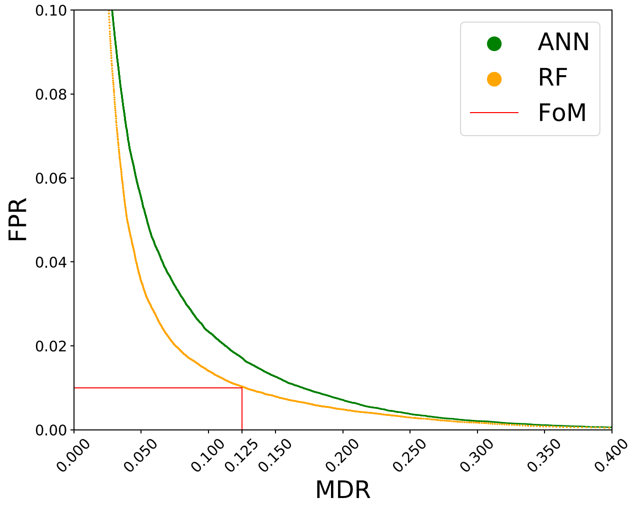

Figure 2. The receiver operator characteristic (ROC) curves of the injection test applied to different learning algorithms. The ANN and RF classifiers are represented by green and orange lines, respectively. The RF classifier shows a better performance, with FoM, indicated by the red line of 12.5 per cent.

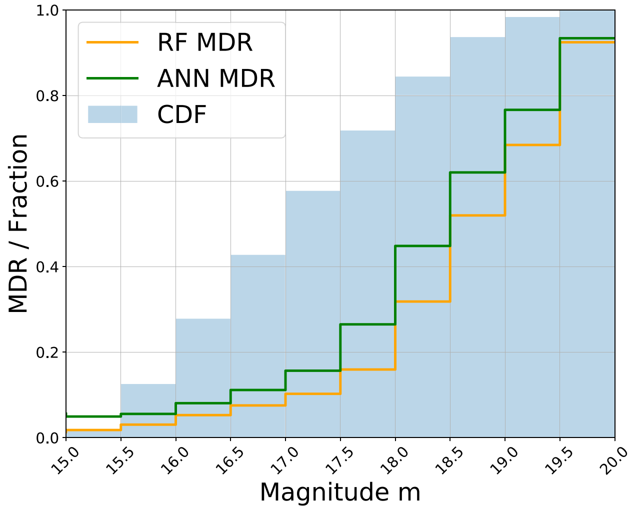

Figure 3. MDR for the injections as a function of magnitude. The RF classifier indicated by the orange line always shows a lower MDR over the ANN model (green line). The blue blocks show the cumulative distribution function (CDF) of the magnitude of the injections.

Table 3. Decision boundaries and prediction accuracies at FPR = 0.01 in the injection test.

Algorithms	Decision boundary	Real accuracy (per cent)	Bogus accuracy (per cent)	FoM (per cent)
RF	0.75	87.5	99.0	12.5
ANN	0.91	83.9	99.0	16.1

We investigate how the MDR varies with the brightness of the detections in Fig. 3. The decision boundary is set to FPR = 0.01 for each of the learning algorithms (see Table 3). The RF classifier has the lowest MDR over the range of magnitudes from m = 15 to 20.

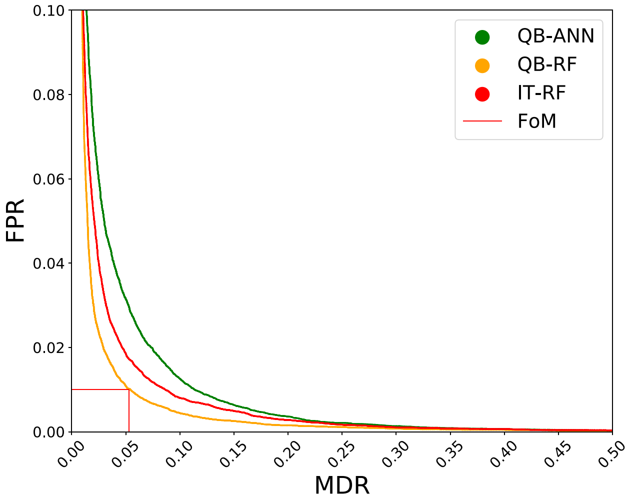

Figure 4. The ROC curves of different learning algorithms tested on the MP test set. QB-ANN and QB-RF classifier are represented by green and orange lines, respectively. The QB-RF classifier shows the best performance with the FoM of 5.2 per cent. The IT-RF classifier represented by the red line shows a consistent performance with the QB-RF classifier.

Table 4. Decision boundaries and prediction accuracies at FPR = 0.01 testing on the MP test set.

Algorithms	Decision boundary	Real accuracy (per cent)	Bogus accuracy (per cent)	FoM (per cent)	F1 score
QB-RF	0.61	94.8	99.0	5.2	0.97
IT-RF	0.55	91.9	99.0	8.1	0.95
QB-ANN	0.86	89.2	99.0	10.8	0.94

We also plot the cumulative distribution function (CDF) against the magnitude in Fig. 3. The constant step size of about 10–15 per cent from m = 16 to 18.5 in the CDF shows a uniform magnitude distribution of the injections in our data set. The decrease in the step size beyond m = 18.5 is due to the drop of the SExtractor recovery rate with the increase of magnitude as we are nearing the limiting magnitude of GOTO.

6.2 Performance on the MP data set

In this section, we include one more RF classifier trained on the injection set (IT-RF) in our analyses. The purpose of comparing with the IT-RF classifier is to show that the classifiers trained on our quick-build training set also perform a consistently with the classifier trained on the data solely collected from the difference images.

We test our classifiers on real data by using our MP test set (see Section 3.3). According to the ROC curves in Fig. 4, the RF classifier trained on the quick-build training set (QB-RF) shows the lowest FoM of 5.2 per cent. Both AUC and FoM also show that QB-RF and IT-RF perform consistently with each other. The decision boundaries used in this section and the FoMs of all classifiers are also showed in Table 4. Since our MP test set is slightly unbalanced with real-to-bogus ratio of 1: 1.3, we also list F1 scores, which helps to estimate the goodness of balance between the recall and the precision, in Table 4. The F1 score of QB-RF is closest to 1, indicating that this model is superior to the other models considered. We also show the confusion matrices for different classifiers at a fixed FPR of

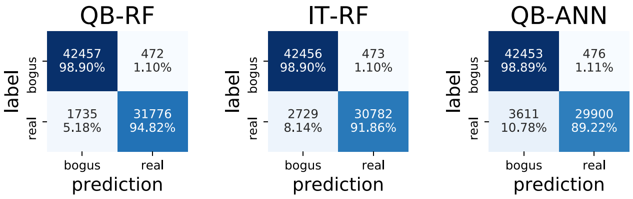

Figure 5. Confusion matrices of different models performing on the MP test set. The decision boundary of each classifier is set at FPR = 0.01. The QB-RF shows the highest real prediction accuracy of 94.8 per cent.

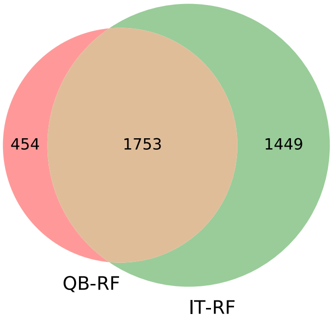

Figure 6. Venn diagram of the number of misclassified sources for QB-RF and IT-RF. It shows that ≈ 80 per cent of the QB-RF misclassifications are also misclassified by IT-RF.

1 per cent in Fig. 5. We plot the Venn diagram in Fig. 6 to compare the misclassification consistency between QB-RF and IT-RF. The intersection is about 80 per cent of the QB-RF population, which implies the misclassifications of the two models are consistent with each other.

Fig. 7 shows that both QB-RF and IT-RF classifiers can separate bogus and real detections in the MP test set effectively. There are small overlapping regions at around 0.5 for the QB-RF distribution and 0.4 for the IT-RF distribution. The difference between these two decision boundaries is caused by the ratio difference between the numbers of the real and the bogus samples in the training sets.

We can see that the results showed in Figs 4 and 8 are different from what the Figs 2 and 3 present. The QB-RF shows a much lower FoM of about 5 per cent in Fig. 4 than in Fig. 2. However, the conclusions that can be drawn from both ROC curves are the same, the QB-RF performs the best in terms of both FoM and AUC. The MDR-mag plots, in Fig. 8, show that the MDR of the QB-RF always stays below 0.3 even up to $m = 20$, which is much lower than the one of >0.9 in Fig. 3. There are several potential factors causing these differences. First, we can see the CDFs in Figs 3 and 8 are different, indicating two different brightness distributions of the real samples in the data sets. In the injection data set, we inject sources with a uniform brightness distribution. On the other hand, in the MP test set, there is a more accurate representation of the generalized magnitude distribution in comparison to the artificial one from our injection set. Secondly, we only use the images taken in a particular field with particular instruments, UT3 and UT4, to build our injection set. In contract, the MP test set includes detections from images taken with a wider range of conditions, with different UTs, fields, image quality scores, etc. Finally, the PSF models used to generate the injections can never be fully representative of the range of genuine detections appearing on the difference images as they are discretized on the image.

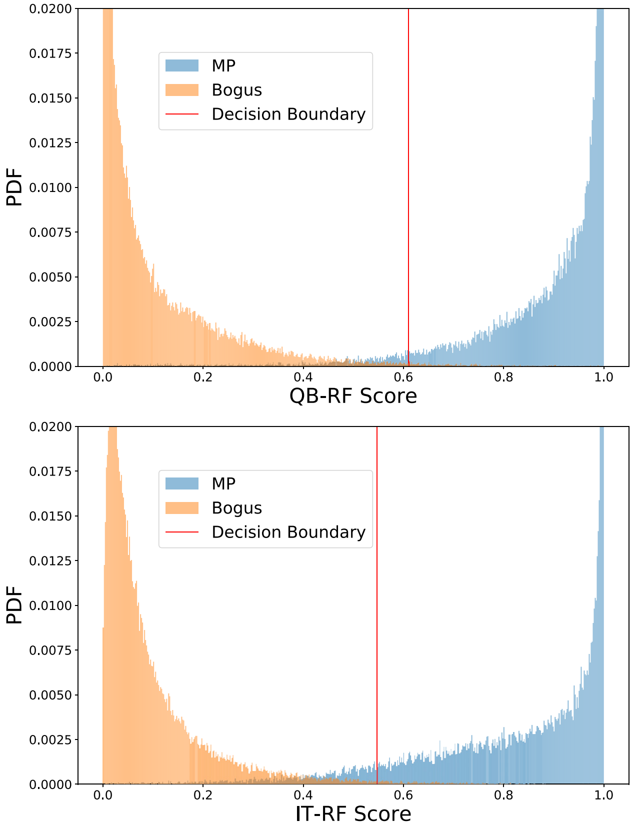

Figure 7. The classification score distributions of the MP test set. The top and bottom panels represent the QB-RF and the IT-RF classifiers respectively. The orange histograms represent the score distribution of the bogus detections, meanwhile the blue histograms represent the distribution of the MPs. The red lines indicate the decision boundaries set at FPR = 0.01.

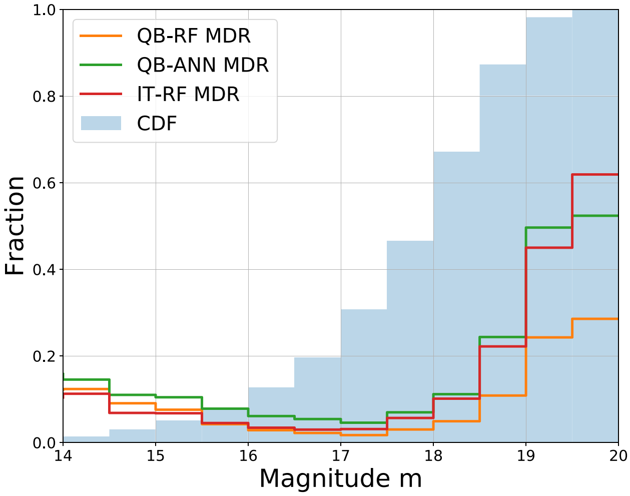

Figure 8. MDR for the MPs as a function of magnitudes. The QB-RF classifier indicated by the orange line always shows the lowest MDR. The MDRs of QB-ANN and IT-RF are also plotted with green and red lines, respectively. The blue blocks show the CDF with magnitude of the MPs.

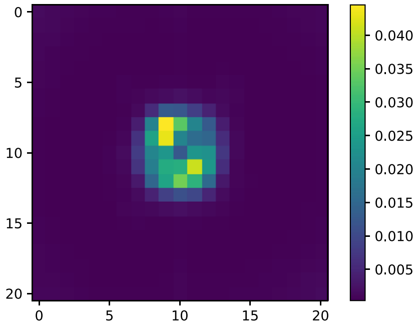

Figure 9. The RF feature importance of each pixel over the stamp. It shows that the central 7 × 7 pixels are the most important features for separating real and bogus detections using our QB-RF classifier.

We provide evidence that the training set constructed using our quick-building strategy is not only fast and convenient, but shows nearly identical performance to the classifiers trained in the traditional way. Since the main scope of this paper is to show how to address the problem of assembling a sufficiently large data set for supervised machine learning, the performance comparison between the different learning algorithms is for reference only. The results might depend on the architectures of the classifiers, the feature representation, etc.

6.3 Feature importance

To understand how the RF classifier calculates the classification score for a detection thumbnail, we can simply plot out the feature importance of each pixel (see Fig. 9). As we expect, to classify whether a detection is real or bogus, the classifier only considers the central 7 × 7 pixels as the most important features. This area is consistent with the 90 per cent percentile of the FWHM distribution, which is 7.8 pixels for GOTO prototype performance, for the real samples in our QB training set.

Fig. 9 shows that the central pixel is not the most important pixel feature among the entire stamp. This could be due to the elongation of the PSF of the real detections we used to train our classifier (see Fig. 1).

Additionally, the pixels outside the 7 × 7 central region have very low values of feature importance. There are two conclusions that can be drawn from this observation. The classification scores for those transients close to bright objects or galaxies would not be affected. However, the subtraction residuals from the bright objects could easily be scored with a high value. Fortunately, the subtraction residuals due to the bright objects can easily be filtered by human vetting which should always be done as a confirmation of the candidates after the automatic real-bogus classification process. Another method of solving this problem is to reject candidates within a certain angular distance from bright objects.

The feature importance of our RF model prompted us to train another classifier with a different stamp size. We used stamp sizes of 7 × 7 and 15 × 15 pixels to train additional models (called QB-RF7 and QB-RF15). Since we use the median pixel value on each stamp as the noise level to perform scaling and filling the masked pixels, if the stamp size is close to the PSF area, the median pixel value may not well represent the noise level. Therefore, we use the original 21 × 21 pixel stamp to obtain the noise level, and then use another crop to generate a smaller pixel stamp for our training features.

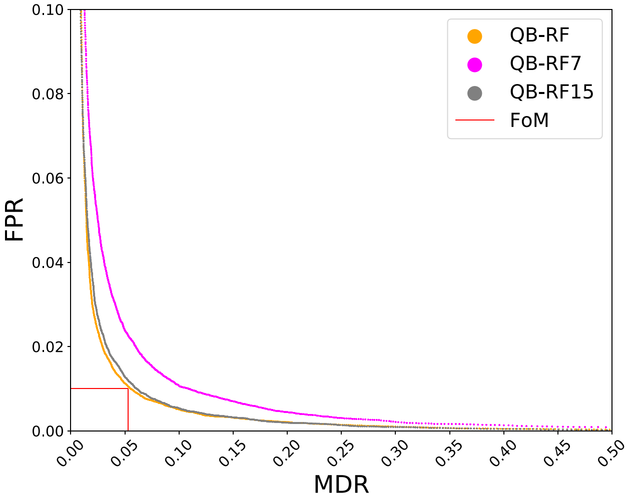

Figure 10. The ROC curves of the QB-RF (orange), the QB-RF7 (magenta), and the QB-RF15 (grey) classifiers tested on the MP test set. The QB-RF7 doubles the FoM compared to the QB-RF. Both QB-RF and QB-RF15 perform consistently.

We use the MP test set only to compare the differences between models trained with different stamp sizes. Fig. 10 shows that the FoM of the QB-RF7 is about 10 per cent, which is about twice that for QB-RF, but the ROC curve of the QB-RF15 is consistent with the QB-RF classifier. Therefore, we suggest using stamp sizes of at least twice the 90 per cent percentile of the FWHM for training.

7 CONCLUSION AND SUMMARY

In this paper, we design and test methods to separate real detections in optical difference imaging from bogus ones, by using machine learning methods. Manually building a large training set is very time consuming that motivates the use of detections in the science images, which should look identical to transients in the subtracted images, as the real sample. Our training set consists of 400 000 real and bogus detections, respectively. We use scaled pixel values over a 21 × 21-pixel stamp centred at the detection position to represent the features of each detection to calculate the real-bogus score.

The RF classifier is shown to have a better performance compared to ANN by testing with the MP data set. We obtain an overall accuracy of 97.1 per cent and FoM of 5.2 per cent with the decision boundary set to 0.61. We also show that the classifier trained on our quick-build training set has a similar performance with the classifier trained on our injection data set.

Compared to the traditional methods used to build a training set for supervised machine learning methods, our strategy can help to build a training set of reasonable size within few days without having to spend weeks to months on manual inspection and human verification. We also show that the performance of the classifier built based on this strategy is comparable to the classifier built by traditional methods.

We also build two other RF classifiers by training on 7 × 7 and 15 × 15 pixel stamps, to study how the performance varies with stamp size. We show that a 15 × 15 pixel stamp is sufficient to train our model. Therefore, we recommend using at least twice the 90 per cent percentile FWHM as the training stamp size.

While the quick-build strategy we use to build our training set is both fast and effective to train our classifier, we do not prescribe this technique to assess the best method of building a classifier overall.

Instead, we suggest it could serve as a preliminary classifier for transient searches with newly-operational optical telescopes, or being ideal for small research collaborations that decide to pursue transient search projects. Since we only use the pixel intensity for performing classification, the idea of this work, in principle, should be directly applicable with other instruments.

ACKNOWLEDGEMENTS

We thank the referee for the comments that improved this paper. The Gravitational-wave Optical Transient Observer (GOTO) project acknowledges the support of the Monash-Warwick Alliance, Warwick University, Monash University, Sheffield University, University of Leicester, Armagh Observatory and Planetarium, the National Astronomical Research Institute of Thailand (NARIT), the Instituto de Astrofísica de Canarias (IAC), and the University of Turku. RB, MK and DMS acknowledge support from the ERC under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 715051; Spiders).

DATA AVAILABILITY

Data products will be available as part of planned GOTO public data releases.

REFERENCES

Abadi M. et al., 2016, TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, OSDI'16: Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation, 265-283
Abbott B. P. et al., 2016, Phys. Rev. Lett., 116, 241102
Abbott B. P. et al., 2017, Phys. Rev. Lett., 119, 161101
Becker A., 2015, Astrophysics Source Code Library, record ascl:1504.004
Berthier J., Vachier F., Thuillot W., Fernique P., Ochsenbein F., Genova F., Lainey V., Arlot J. E., 2006, in Gabriel C., Arviset C., Ponz D., Enrique S., eds, ASP Conf. Ser. Vol. 351, Astronomical Data Analysis Software and Systems XV. Astron. Soc. Pac., San Francisco, p. 367
Bertin E., 2000, Mining the Sky , Proceedings of the MPA/ESO/MPE Workshop Held at Garching, Germany
Bertin E., Arnouts S., 1996, A&AS, 117, 393
Bloom J. S. et al., 2012, PASP, 124, 1175
Breiman L., 2001, Mach. Learn., 45, 5
Brink H., Richards J. W., Poznanski D., Bloom J. S., Rice J., Negahban S., Wainwright M., 2013, MNRAS, 435, 1047
Cabrera-Vives G., Reyes I., F"orster F., Estévez P. A., Maureira J., 2016, 2016 International Joint Conference on Neural Networks (IJCNN), p. 251
Cabrera-Vives G., Reyes I., Förster F., Estévez P. A., Maureira J.-C., 2017, ApJ, 836, 97
Coulter D. A. et al., 2017, Science, 358, 1556
Drake A. J. et al., 2009, ApJ, 696, 870
Dyer M. J., Dhillon V. S., Littlefair S., Steeghs D., Ulaczyk K., Chote P., Galloway D., Rol E., 2018, Observatory Operations: Strategies, Processes, and Systems VII, Vol. 107040C, SPIE Proceedings
Frieman J. A. et al., 2008, AJ, 135, 338
Gieseke F. et al., 2017, MNRAS, 472, 3101 (G1)
Heinze A. N. et al., 2018, AJ, 156, 241
Ivezić Ž. et al., 2019, ApJ, 873, 111
Kaiser N. et al., 2010, in Stepp L. M., Gilmozzi R., Hall H. J., eds, Proc. SPIE Cnf. Ser. Vol. 7730, Ground-Based and Airborne Telescopes III. SPIE, Bellingham, p. 77330e
McCulloch W. S., Pitts W., 1943, Bull. Math. Biophys., 5, 115
Masci F. J. et al., 2018, PASP, 131, 018003
Metzger B. D., 2017, Living Rev. Relativ., 20, 3
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Tody D., 1986, The IRAF Data Reduction and Analysis System, Proc. SPIE 0627, Instrumentation in Astronomy VI
Tody D., 1993, IRAF in the Nineties, Vol. 52, Astronomical Data Analysis Software and Systems II, A.S.P. Conference Series
Tonry J. L. et al., 2018, PASP, 130, 064505
Wolf C. et al., 2018, Publ. Astron. Soc. Aust., 35, e010
Wright D. E. et al., 2015, MNRAS, 449, 451 (W15)

This paper has been typeset from a TEX/LATEX file prepared by the author.