Classification of multiwavelength transients with machine learning

K. Sooknunan, M. Lochner, Bruce A. Bassett, H. V. Peiris, R. Fender, A. J. Stewart, M. Pietka, P. A. Woudt, J. D. McEwen and O. Lahav

1Department of Maths and Applied Maths, University of Cape Town, Cape Town 7935, South Africa
2Astrophysics Group, Imperial College London, Blackett Laboratory, Prince Consort Road, London SW7 2AZ, UK
3Department of Physics and Astronomy, University of the Western Cape, Bellville, Cape Town 7535, South Africa
4South African Radio Astronomy Observatory (SARAO), The Park, Park Road, Pinelands, Cape Town 7405, South Africa
5African Institute for Mathematical Sciences, 6 Melrose Road, Muizenberg 7945, South Africa
6South African Astronomical Observatory, Observatory, Cape Town 7925, South Africa
7Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UK
8The Oskar Klein Centre for Cosmoparticle Physics, Department of Physics, Stockholm University, AlbaNova, SE-10691 Stockholm, Sweden
9Astrophysics, Department of Physics of Oxford, Keble Road, Oxford OX1 3RH, UK
10Department of Astronomy, University of Cape Town, Cape Town 7701, South Africa
11Sydney Institute for Astronomy, School of Physics, The University of Sydney, NSW 2006, Australia
12Mullard Space Science Laboratory, University College London, Surrey RH5 6NT, UK

Accepted 2020 December 8. Received 2020 December 8; in original form 2020 June 2

ABSTRACT

With the advent of powerful telescopes such as the Square Kilometer Array and the Vera C. Rubin Observatory, we are entering an era of multiwavelength transient astronomy that will lead to a dramatic increase in data volume. Machine learning techniques are well suited to address this data challenge and rapidly classify newly detected transients. We present a multiwavelength classification algorithm consisting of three steps: (1) interpolation and augmentation of the data using Gaussian processes; (2) feature extraction using wavelets; and (3) classification with random forests. Augmentation provides improved performance at test time by balancing the classes and adding diversity into the training set. In the first application of machine learning to the classification of real radio transient data, we apply our technique to the Green Bank Interferometer and other radio light curves. We find we are able to accurately classify most of the 11 classes of radio variables and transients after just eight hours of observations, achieving an overall test accuracy of 78 per cent. We fully investigate the impact of the small sample size of 82 publicly available light curves and use data augmentation techniques to mitigate the effect. We also show that on a significantly larger simulated representative training set that the algorithm achieves an overall accuracy of 97 per cent, illustrating that the method is likely to provide excellent performance on future surveys. Finally, we demonstrate the effectiveness of simultaneous multiwavelength observations by showing how incorporating just one optical data point into the analysis improves the accuracy of the worst performing class by 19 per cent.

Key words: methods: data analysis.

1 INTRODUCTION

In the coming years, radio astronomy will enter a new era of deep field surveys with new telescopes such as the Square Kilometre Array (SKA) and its precursors, MeerKAT and the Australian Square Kilometre Array Pathfinder (ASKAP). These telescopes will achieve unprecedented sensitivity and resolution. Large science projects such as ThunderKAT (Armstrong et al. 2018) on MeerKAT will dramatically increase the detected number of radio transients. In the past, radio transient data sets have been small, allowing spectroscopic classification of all objects of interest. As the event rate increases, follow-up resources must be prioritized by making use of early classification of the radio data. Machine learning algorithms have proven themselves invaluable in this context (Ball & Brunner 2010).

There has been a substantial amount of work done with machine learning in astronomy over the last decade. This includes research done by Bailner-Jones (2002) in stellar classification, image-based classification of supernovae (SNe; Romano, Aragon & Ding 2006; Bailey et al. 2007), and classifying variable stars (Richards et al. 2011). In recent years, these algorithms have been used successfully in classifying optical transients, such as classification of transients in SDSS images (du Buisson et al. 2015), SNe (e.g. Newling et al. 2011), Ishida & de Souza (2013), Karpenka, Feroz & Hobson (2013), Lochner et al. (2016), Möller & de Boissière (2019), and classifying variable sources (Farrell, Murphy & Lo 2015), or general optical transients (Mahabal et al. 2017). Some machine learning methods have been investigated for the upcoming ASKAP survey for variables and slow transients (VAST; Rebbapragada et al. 2012; Murphy et al. 2013), but these were only applied to optical data.

* E-mail: k.sooknunan19@imperial.ac.uk
1www.skatelescope.org
2www.ska.ac.za/science-engineering/meerkat
3www.atnf.csiro.au/projects/askap/index.html

© 2020 The Author(s)
Published by Oxford University Press on behalf of Royal Astronomical Society
In the burgeoning era of multimessenger astronomy, incorporating data from different telescopes could dramatically improve classification of events. A prime example of this is the MeerLICHT\(^4\) telescope (Bloemen et al. 2016), an optical telescope whose observing schedule is synchronized with that of the (night time) observations of the radio telescope MeerKAT, resulting in simultaneous optical and radio observations of transients. Indeed, a primary motivation for undertaking this work now is that MeerKAT has recently begun to take scientific data and it is anticipated to detect hundreds of transients over the next few years. Alert streams from telescopes such as Fermi\(^5\) and the Vera C. Rubin Observatory\(^6\) (LSST Science Collaboration 2009) will also enable rapid coordination for multimessenger observations. Combining these data sources necessitates a new general framework for multimessenger machine learning.

In this paper, we outline a method for the automatic classification of radio transients, based on Lochner et al. (2016), that makes use of multiwavelength data and machine learning. We define how to incorporate data in other wavelengths and alert streams from other telescopes. The techniques proposed by Lochner et al. (2016) were only used on optical SNe. Here, we apply this technique to radio data and show that it can also be effective for objects that vary on dramatically different time-scales; SNe and flare stars (FSs), for instance, vary on the time-scales of seconds, while active galactic nucleus (AGN) can vary on the time-scale of months.

Due to the limited quantity of available radio transient light curves, we develop a data augmentation technique to increase the sample size and improve classification performance. Our data augmentation technique achieves three things: it breaks up long light curves into smaller samples increasing the size of the training set, it balances the number of objects in each class that reduces bias in the machine learning and it adds a small amount of additional variance to the training data that is statistically consistent with the original light curve.

We test our method on existing radio transient light curves, exploring the effects of non-representative training sets (i.e. when the algorithm is tested on objects that are dissimilar to those in the training set). We also demonstrate, with an example, the effect of including optical data to improve classification accuracy.

2 BACKGROUND

2.1 Radio transients

A transient is an astronomical object observed to have a time-dependent brightness (flux). The time-scales of variability range from milliseconds to a few years. Radio transients have emission frequencies in the radio regime (for a review on radio transients, see Fender & Bell 2011). Transient events are typically divided into two kinds: incoherent synchrotron events and coherent burst events.

Incoherent synchrotron events are thought to be caused by a high-energy phenomenon, from which a large amount of energy is released over a longer time-scale (of the order of minutes or larger). When in a steady state, this energy release is limited to a brightness temperature of \(T \leq 10^{22}\) K (Fender et al. 2015).

Coherent bursts occur on much shorter time-scales (of the order of seconds) and can have brightness temperatures up to \(10^{30}\) K. This type of emission can only be observed using a special mode on radio telescopes (Fender et al. 2015). For this study, therefore, we will only consider incoherent synchrotron transient events. We assume that a light curve is generated from measuring the fluxes from radio images. We include the following transients in our study: AGNs, algols, FSs, gamma-ray bursts (GRBs), kilonovae, magnetars, novae, RS canum venticorums (RSCVn), SNe, tidal disruption events (TDEs), and X-ray binaries (XRBs).

Looking at the raw radio light curves of transients, while some objects exhibit obvious differences (such as binary star systems and AGNs) others look more similar. Contextual information (such as the location of the object) can be invaluable in telling the difference between classes, as can multiwavelength data. We begin by studying how well we can distinguish between classes using radio data alone, and then show how including contextual and multiwavelength information can improve the classification accuracy.

2.2 Machine learning overview

Machine learning can broadly be split into two approaches: supervised and unsupervised learning. In this study, we use supervised learning. A supervised machine learning algorithm automatically learns a model given a set of known inputs and outputs, called a training set. The now trained algorithm can be given new inputs, called a test set, that it will then map to an output. In terms of a classification problem, the inputs constitute objects to be classified, and the outputs are the class labels assigned to each object. For an in depth review of machine learning, see Mitchell (1997) or MacKay (2003).

In this study, we use the ‘random forests’ algorithm (Ho 1995; Breiman 2001). Random forests have been shown to outperform other algorithms in a variety of cases (Caruana & Niculescu-Mizil 2006; Liu et al. 2013; Lochner et al. 2016). Deep learning algorithms\(^7\) such as LSTMs\(^8\) (Hochreiter & Schmidhuber 1997) would be an interesting approach to this problem. However, we do not consider them here due to the requirement of large training sets and significant computational resources, especially in light of the excellent performance of faster traditional techniques.

2.2.1 Random forests

Ensemble methods like random forests (Ho 1995; Breiman 2001) build robust classifiers out of a multitude of weak learners such as decision trees. A decision tree creates a mapping, by making a series of ‘yes/no’ decisions, from an input vector (the feature vector) to an output label (the class) (Ball & Brunner 2010). The algorithm creates this mapping by making decisions based on whether or not a given component of the feature vector falls into some range. One of the main drawbacks of decision trees is the high variance on the labels it outputs.

This problem can be overcome by training many separate trees and taking the average of the output. Random forests perform an additional step, each tree in the ‘forest’ is trained on a random subset of the total feature set. This leads to more robust overall predictions. We used the package scikit learn\(^9\) (Pedregosa et al. 2011) to implement the random forest classifier.

\(^4\)www.meerlicht.uct.ac.za
\(^5\)https://fermi.gsfc.nasa.gov
\(^6\)www.lsst.org
\(^7\)For a review on deep learning algorithms, see Vargas, Mosavi & Ruiz (2017).
\(^8\)An LSTM or long short term memory network is a type of deep learning algorithm that is primarily used when dealing with time series data.
\(^9\)www.scikit-learn.org
2.2.2 Feature extraction in machine learning

Classical machine learning techniques can seldom use data in its raw format for classification. Feature extraction is a technique used to reduce the dimensionality of the data by summarizing the information contained in the original data. The features one uses should also be well-separated between classes. Taking a repeating light curve as an example, features one could extract are the frequency, the amplitude, the phase, etc. For more on feature extraction, see Li et al. (2016). An obvious choice of simple features for this problem would be changes of flux values over specific time periods. However, we found these were inadequate to capture the variation between classes (see Section B1) and so instead follow the feature extraction procedure used in Lochner et al. (2016). In Section 3.2.2, we outline the wavelet decomposition approach used, which resulted in much higher performance.

2.2.3 Visualizing features

Visualizing feature vectors is quite difficult because of their often high-dimensional nature. In this paper, we use two methods for feature visualization.

- \textit{t-distributed stochastic neighbour embedding:} One tool commonly used to visualize higher-dimensional spaces is \textit{t}-distributed stochastic neighbour embedding or \textit{t-SNE} (van der Maaten & Hinton 2008). It works by computing the probability that two points are similar in the higher dimensional space based on its Euclidean distance. It does this for every pair of points in the feature set, then attempts to find a lower dimensional representation of these points that preserves the probability distribution. Thus, points clustered in this lower dimensional representation correspond to points clustered in the original higher dimensional space. \textit{t-SNE} uses Student’s \textit{t}-distribution when determining the degree of similarity between points. It is important to note that distances are not preserved when transforming to the lower dimensional space but rather the probability distributions. Therefore, while these visualizations can be used to give us some intuition into how the data points are distributed in higher dimensional space, they cannot be used to infer the classes of the data points. We stress that \textit{t-SNE} plots are useful tools for visualization purposes only, and cannot be used as classifiers themselves due to their stochastic nature.

- \textit{Uniform manifold approximation and projection for dimension reduction:} Recently, a new visualization and dimensionality reduction technique uniform manifold approximation and projection for dimension reduction (UMAP) from McInnes et al. (2018) has begun to overtake \textit{t-SNE} in popularity, due to its superior computational performance and interpretability.

We make use of both techniques to visualize our feature space.

2.2.4 Training, testing, and cross validation

In machine learning, a data set is generally split into two main subsets: the training set and the test set. The algorithm is given the training set from which to learn the parameters of the model. The test set is reserved in order to check if the algorithm has learned an accurate model, and is only presented to the algorithm after the training step is complete. In some cases, an algorithm can perform very well on the training set but perform poorly on the test set. This occurs when the algorithm overfits the model parameters to the test set and hence the model will not generalize to the training set. One method for overcoming this is known as \textit{k}-fold cross validation. Instead of splitting the data set into two subsets, the data set is split into \( k > 2 \) subsets. One of these subsets is then used as the validation set while the others are used as the training set. This is then repeated \( k \) times until all \( k \) subsets have been used as a test set and average results are used to evaluate the algorithm.

Machine learning models have two types of parameters. The first are the parameters that are learned by the algorithm during training as mentioned above. The second type is known as hyperparameters. These are parameters that are not learned during training but are set by the user. These parameters can also be optimized by specifying a range for each hyperparameter, searching through this hyperparameter space, and using cross-validation to choose the hyperparameters with the best performance. Three-fold cross-validation was used to optimize the random forest algorithm hyperparameters.

2.2.5 Data augmentation

Data augmentation is widely used in machine learning when dealing with small data sets (Hoyle et al. 2015; Oviedo et al. 2019). There are three main benefits of data augmentation: the first is the creation of larger training sets which is important for complex algorithms with many parameters (such as deep neural networks); the second benefit is that it helps avoid overfitting by introducing variation in the original data set; finally it allows the balancing of classes which may otherwise cause biases in the classifier. Augmentation is used extensively in the field of deep learning for image classification: training images are rotated, cropped, translated, and zoomed to create new examples that can dramatically improve final performance (e.g. Kim & Brunner 2016; Perez & Wang 2017). The technique synthetic minority oversampling technique is commonly used to balance class numbers and add variance by simulating new data in feature space (Chawla et al. 2011). Naul et al. (2018) used data augmentation to help train a recurrent neural network in the classification of variable stars. Recently, Revsbech, Trotta & van Dyk (2018) and Boone (2019) have made use of Gaussian processes (GPs) to augment non-representative training data to increase accuracy for classification of SNe and other optical transients.

2.2.6 Evaluating machine learning results

The performance of a machine learning algorithm can be measured in different ways. We will evaluate our results using confusion matrices. A confusion matrix is a plot showing the true label of the object on one axis and the label predicted by the machine learning algorithm on the other. For the simplest classification problem, a binary classification problem, the confusion matrix would be a \( 2 \times 2 \) matrix with the true positives and true negatives along the diagonal, and the false negatives and false positives on the off-diagonals. Therefore, confusion matrices show how well the algorithm classifies each class. Classes classified correctly would appear on the diagonal, and incorrect classifications would appear on the off-diagonals.

3 GENERAL APPROACH TO MULTIWAVELENGTH TRANSIENT CLASSIFICATION

Drawing heavily from Lochner et al. (2016), we outline a general approach to classifying transients with multimessenger data. The approach is split into two main sections; the first deals with combining data from different sources and the second builds a machine learning classifier that uses light-curve data of any wavelength. This creates a general approach useful for combining all sources of information that may be useful to classifying transients, in addition to classification using the light curves themselves.
3.1 Combining multiple data sources

The data used to classify a source need not be information extracted directly from the light curves. It can also be prior or external information about the sources, such as fluxes of the source in different wavelengths or, contextual information, such as position of the object in the sky. Information from alert streams from other observatories can also be added as external information (e.g. the presence of gamma-ray emission or a gravitational wave detected by LIGO\textsuperscript{10} in the region of a new radio transient source can be highly discriminating).

There are two methods for incorporating information from other sources:

(i) **Probabilistic approach**: most machine learning classification algorithms are capable of producing a score that can be interpreted as a probability of an object belonging to a particular class. To combine this with external information, such as the presence of a coincident alert at another wavelength, we can calculate the prior probability, \( P(C) \), of the object being in a certain class \( C \), given all prior information. This probability, \( P(C) \), would then be multiplied by the probability given by the classifier to give a final probability of some object being in class \( C \).

(ii) **Extra features**: the second method is to use the information as an extra feature in the machine learning process. For example, if one has a flux measurement at any other wavelength, one could add that flux as a feature. The advantage of this approach is that correlations between the different features are learned automatically by the machine learning algorithm, potentially resulting in improved classification accuracy.

The disadvantage of this latter approach is that most machine learning algorithms do not intrinsically deal well with missing data. This could happen if, for instance, MeerKAT detects a transient during a daytime observation when MeerLICHT cannot observe. While feature imputation techniques exist (Quinlan 1993), a more interpretable approach may be to combine the probabilities where, if data are missing, a default probability based on prior observations (for example, known transient rates) can be used.

The specific set-up of the problem will dictate which approach is more appropriate, but formalizing this process is a step towards automated multimessenger machine learning pipelines, that can then be fed into downstream analysis including spectroscopic follow-up prioritization.

3.2 General approach to transient classification with light curves

Our method for transient classification follows almost identically the technique outlined in Lochner et al. (2016), which was used for classification of SNe light curves at optical wavelengths. The technique is applicable to any transient (or indeed, almost any time series data).

3.2.1 Step 1: Interpolation

Given some light curve data, \( D \), the first step is to interpolate the data so that it is on a uniform grid. This is done using GPs (Rasmussen & Williams 2005), since the mean function derived from a GP is an extremely robust interpolator in the presence of noisy data. We use GPs as described in Section 4.2 for interpolation.

3.2.2 Step 2: Wavelet decomposition

Time series data can be decomposed into a linear combination of basis functions

\[
 f(x) = \sum k a_k \phi_k(x),
\]

where \( \phi_k(x) \) are orthogonal basis functions and \( a_k \) are the respective coefficients.

This is a common approach in signal processing and can be a powerful tool for feature extraction, to obtain the set of coefficients used as the features with a machine learning algorithm. One widely used form of this is a Fourier decomposition, where a signal can be decomposed into the component frequencies. However, the Fourier decomposition loses all localization information and is thus mostly applicable to regular, repeating signals.

By contrast, in transient classification, the object may be observed at any point in its light curve and the algorithm must determine its class in this setting. Thus, we require a decomposition method that is translation-invariant but still sensitive to the intrinsic shape of the curve. A form of decomposition that is approximately scale and translation-invariant is known as the stationary wavelet transform (Holschneider et al. 1989; Mallat 2009). Following its successful use in Lochner et al. (2016) and Narayan et al. (2018), we make use of the stationary wavelet transform with the `syrlet` family, as implemented in the package PYWAVELETS.\textsuperscript{11}

3.2.3 Step 3: Dimensionality reduction with PCA

The stationary wavelet transform produces a large number of redundant features, too many for standard machine learning techniques. Therefore, we need to reduce the number of features while keeping as much information as possible contained in the full feature set. Principal component analysis (PCA; Pearson 1901) is a dimensionality reduction technique. It is a linear transformation that decorrelates a set of correlated variables by calculating the covariance matrix of the data set. The eigenvalues and eigenvectors of this matrix are computed. The total variance of the data can be quantified by calculating the sum total of the eigenvalues. The variance is a proxy for the information contained in the data set. Thus, by maximizing the variance of the reduced data set, the majority of information is conserved. The eigenvectors with the largest eigenvalues, which describe the majority of the variability in the data set are stored; smaller eigenvalues are disregarded. The number of eigenvectors kept are decided by the fraction of variability one would want to retain in the data set; variability is defined as the sum of all eigenvalues. For example, if we want to keep 90 per cent variability in our data set, then we would keep the corresponding eigenvectors of the largest eigenvalues (in descending value) until their sum equals 90 per cent of the sum of all eigenvalues. The large number of coefficients that we obtain from PYWAVELETS can be projected on to the stored eigenvectors, producing a new set of eigenvalues.

4 APPLICATION TO RADIO TRANSIENTS

Motivated by the expected increase in new transient detections with modern radio telescopes, we apply our general approach to existing radio transient data. Because these data are limited, we use the data interpolation and augmentation technique described in Section 4.2.

\textsuperscript{10}www.ligo.org

\textsuperscript{11}https://github.com/PyWavelets/pywt
Table 1. Breakdown of radio transient data into the relevant types and data sources. GBI refers to data collected from the Green Bank Interferometer. From Lit. refers to data collected from the literature.

| Type     | From Lit. | GBI | Total |
|----------|-----------|-----|-------|
| AGN      | 17        | 13  | 30    |
| Algol    | 1         | 2   | 3     |
| FS       | 5         | 0   | 5     |
| GRB      | 4         | 0   | 4     |
| Kilonova | 1         | 0   | 1     |
| Magnetar | 1         | 0   | 1     |
| Nova     | 8         | 0   | 8     |
| RSCVn    | 0         | 2   | 2     |
| SN       | 13        | 0   | 13    |
| TDE      | 2         | 0   | 2     |
| XRB      | 11        | 9   | 20    |
| Total    | 63        | 26  | 89    |

as a statistically robust method to simulate light curves for training and testing. We follow the feature extraction method described in Section 3 and also illustrate the effect of incorporating additional data by including contextual information and optical data.

4.1 Radio data

The radio transient data used were collected by Pietka et al. (2015) except for the kilonova light curve, which was collected by Dobie et al. (2018).

Most of the data are obtained from the Green Bank Interferometer\(^{12}\) (GBI). All the radio observations were done at frequencies between 5 and 8 GHz. Data collected from the GBI have a much higher cadence than the data obtained from the literature, which makes is the longest observation time for which we have measurements of this kind. We restrict our study to a time-scale of 8 h, which is the longest observation time for which we have measurements for all classes (see Fig. 2). Classification on this time-scale will also allow relatively prompt follow-up triggers. The technique is applicable on even shorter time-scales, even if there are very few flux measurements, although classification accuracy will likely decrease.

4.2 Interpolation

The feature extraction method described in Section 3.2.2 requires the input data to be sampled on a regular grid. The data used in this work all have very different cadences hence the light curves need to be interpolated. As mentioned in Section 3.2.1, we use GPs\(^{13}\) (Rasmussen & Williams 2005) to interpolate the light curves due to its robustness in the presence of noisy data.

GP regression on each light curve was performed using each of these combinations $K_1, K_2, K_3$ and error. It was found that these combinations fit the data well. Due to the current limited data set, it is likely that there may be new light curves that are not well described by these combinations. More studies would need to be done to test this once a larger data set has been acquired. We believe that as long as the GP does not overfit, and fits sufficiently well, then the choice of kernel is unlikely to impact the final classification. This is because another feature extraction step is done first, and so is independent of our kernel choice. However, it could have an impact on the data augmentation and this should be studied for a larger data set.

GP regression on each light curve was performed using each of these combinations $K_1, K_2, K_3$, and $K_4$ were formulated through trial and error. It was found that these combinations fit the data set well. Due to the current limited data set, it is likely that there may be new light curves that are not well described by these combinations. More studies would need to be done to test this once a larger data set has been acquired. We believe that as long as the GP does not overfit, and fits sufficiently well, then the choice of kernel is unlikely to impact the final classification. This is because another feature extraction step is done first, and so is independent of our kernel choice. However, it could have an impact on the data augmentation and this should be studied for a larger data set.

GP regression on each light curve was performed using each of these combinations $K_1, K_2, K_3$, and $K_4$ were formulated through trial and error. It was found that these combinations fit the data set well. Due to the current limited data set, it is likely that there may be new light curves that are not well described by these combinations. More studies would need to be done to test this once a larger data set has been acquired. We believe that as long as the GP does not overfit, and fits sufficiently well, then the choice of kernel is unlikely to impact the final classification. This is because another feature extraction step is done first, and so is independent of our kernel choice. However, it could have an impact on the data augmentation and this should be studied for a larger data set.

GPs provide a stochastic interpolation formalism where the joint distribution of all random variables follow a multivariate normal distribution. Thus, every subset of the GP can be characterized fully by its mean and covariance functions. This allows a GP to generate new data that are consistent with the error bars on the observed data. We use GEORGE (Ambikasaran et al. 2014) to perform our GP regression.

The characteristic time-scales on which light curves vary can be highly class dependant; SNe may vary smoothly over months while FSs may change over an hour. GPs handle such a range of time-scales through the covariance function (also known as kernels), $k(\Delta t)$, which specifies how quickly the light curves flux values became uncorrelated as the temporal separation between two times, $\Delta t$, increases.

We will use covariance functions that are stationary, that is they depend only on the time difference between data, $\Delta t = |t_i - t_j|$, and a characteristic time-scale parameter $\ell$.

In order to get good fits to the wide range of light curves in our data, we empirically found that we needed a combination of three covariance functions, or kernels, to be used for regression, which are defined below.

The first is the Gaussian radial basis function, given by

$$k_{\text{rad}}(\Delta t) = \exp\left(-\frac{\Delta t^2}{2\ell^2}\right).$$

The second kernel we use is the exponential function given by

$$k_{\text{exp}}(\Delta t) = \exp\left(-\frac{\Delta t}{\ell}\right).$$

Lastly, we use the exponential sine-squared function, given by

$$k_{\text{sine}}(\Delta t) = \exp\left(-\Gamma \sin^2\left[\frac{\pi}{\ell} \Delta t\right]\right),$$

where $\Gamma$ is a hyperparameter that is allowed to take any value.

Given the general temporal variations of the different classes, we found empirically that one of the following three combinations of these kernels provides a good fit to the data:

$$K_1 = w_1 k_{\text{rad}}(\Delta t),$$

$$K_2 = w_2 k_{\text{rad}}(\Delta t) + w_3 k_{\text{rad}}(\Delta t) k_{\text{sine}}(\Delta t),$$

$$K_3 = w_4 k_{\text{rad}}(\Delta t) + w_5 k_{\text{exp}}(\Delta t),$$

where the $w_i$ are the weights of each kernel.

The combinations $K_1$, $K_2$, and $K_3$ were formulated through trial and error. It was found that these combinations fit the data set well. Due to the current limited data set, it is likely that there may be new light curves that are not well described by these combinations. More studies would need to be done to test this once a larger data set has been acquired. We believe that as long as the GP does not overfit, and fits sufficiently well, then the choice of kernel is unlikely to impact the final classification. This is because another feature extraction step is done first, and so is independent of our kernel choice. However, it could have an impact on the data augmentation and this should be studied for a larger data set.

\(^{12}\)https://public.nrao.edu/telescopes/green-bank-interferometer

\(^{13}\)Also known as Kriging (Matheron & Blondel 1962).
Figure 1. An example light curve for each of the radio transient types plotted as flux (in Jy) as a function of time (in days). It is clear that the data are taken on extremely different time-scales. The original data points are shown in grey and the mean of the Gaussian process is shown in blue. For the SN light curve, one standard deviation away from the mean is shown as a blue envelope. The black lines shown in the SN light curve are two random draws from the GP. It can be seen that these lines are different from, but still consistent with the original data. This data were collected by Pietka, Fender & Keane (2015).
Figure 2. The number of light curves for each class as a function of the length of observation for that source (in other words, there are \( y \) objects with light curves that were observed for at least time \( x \)). There are clear observational biases here: objects that vary on longer time-scales such as AGNs or SNe are observed for longer periods and there are greater number of them. It can be seen that 8 h is the shortest observation period for which all the classes have a non-zero number of light curves.

BFGS) optimization algorithm from the SCIPY\(^{14}\) package, to obtain the best set of hyperparameters for each of the kernel combinations, \( K_1, K_2, K_3 \). The \( K_i \) with the lowest negative log likelihood was used to construct the GP from which to sample for each light curve.

The advantage of GP regression with a learned covariance kernel is that it provides an average light curve as well as Gaussian estimates for the uncertainties in the light curve at all times, learned from the data of that specific light curve. This is the main advantage over other interpolation methods such as spline, where there is no adaptation to the data.

An example GP for an SN light curve is shown in Fig. 1, along with the \( 1 - \sigma \) uncertainty band arising from the covariance kernel specific to this example.

4.3 Augmentation

While the total number of light curves in the data set is small, it can be seen from Fig. 2 that most of the light curves are longer than 1 month. Thus, one of the ways in which the number of light curves in the data set can be increased is by splitting the light curves into smaller chunks. GPs fit a function to the original light curves which allows for a statistically rigorous method to sample between any two points \( t_1 \) and \( t_2 \) on the light curve provided that \( t_2 - t_1 < T \), where \( T \) is the total length of the original light curve. This creates new shorter light curves. By varying the starting point of the sample, the original light curve can be broken up into smaller chunks.

Because this method uses GPs, it takes into account the uncertainties of the original data set which most common augmentation techniques do not. In addition, we are not restricted to sampling the mean of the GP. We are also able to draw samples from the GP distribution thus creating new light curves that are statistically consistent with the original data. This method of data augmentation amounts to a combination of Revsbech et al. (2018) and Cui, Chen & Chen (2016).

Finally, augmentation is capable of balancing the classes, artificially enforcing equal numbers of examples in each class. Machine learning algorithms are well known to be biased in the face of highly unbalanced classes and tend to label most objects as the dominant class(es). The use of GPs allows a natural way to construct balanced training sets that do not disfavor the rarer classes.

Appendix D includes a rigorous study of the impact of these three aspects of our augmentation technique: the addition of more examples, the balancing of classes, and the increase of variance in the training set.

4.4 Feature extraction

The feature extraction method used is described in Section 3.2.2. First GP regression was performed on the original data set. From this, many example light curves can be generated, each statistically consistent with the original data, which allows us to generate a realistic synthetic data set of any size. First, a reference time, \( t_0 \), was drawn at random to be somewhere within the original light curve. This was done to simulate the fact that the transient may be detected at any point on its light curve. This created many new light curves. By varying the starting point of the sample, the original light curve can be broken up into smaller chunks.

Because this method uses GPs, it takes into account the uncertainties of the original data set which most common augmentation techniques do not. In addition, we are not restricted to sampling the mean of the GP. We are also able to draw samples from the GP distribution thus creating new light curves that are statistically consistent with the original data. This method of data augmentation amounts to a combination of Revsbech et al. (2018) and Cui, Chen & Chen (2016).

Finally, augmentation is capable of balancing of the classes, artificially enforcing equal numbers of examples in each class. Machine learning algorithms are well known to be biased in the face of highly unbalanced classes and tend to label most objects as the dominant class(es). The use of GPs allows a natural way to construct balanced training sets that do not disfavor the rarer classes.

Appendix D includes a rigorous study of the impact of these three aspects of our augmentation technique: the addition of more examples, the balancing of classes, and the increase of variance in the training set.
4.5 Incorporating external information

Some classes can have similar light curves but are generally found in different parts of the sky. For example, novae tend to occur in the Galactic plane, while SNe are more likely to be extragalactic.

In order to break the degeneracies between these classes, we added a feature that characterizes where the object is in the sky. Telescopes will always have access to the position in the sky in which it is pointing, hence the drawback of adding features that may sometimes be missing, as outlined in Section 3.1, will not be an issue.

The RA and Dec. coordinates were obtained for all the objects in our data set using a combination of Simbad\textsuperscript{15} and NED.\textsuperscript{16} These coordinates were then converted to Galactic coordinates. If we were to copy these coordinates to each of the augmented light curves, the ML algorithm could learn a correlation between the coordinates and the original light curves, which would lead to biased results.

Therefore, instead we defined a feature that specified whether or not the object was in the Galactic plane. Any object with a Galactic declination of above 10\textdegree{} or below −10\textdegree{} was considered to be out of the Galactic plane. This new feature (which is either 0 or 1) is then appended to the augmented light-curve features.

5 RESULTS

5.1 The effect of training sets

The radio transient data currently available are unfortunately too small to use as a training set for machine learning. Machine learning algorithms will tend to generalize poorly from small training sets. However, by using the method outlined in Section 4.4 to split up the light curves into smaller chunks and also balance the classes, we are able to build a large enough training set to use for machine learning. While the simulating of new light curves using GPs does add some variability to the training set, it does not adequately model the real world variability between the different classes. Thus, in order to test the performance of the classifier in a more realistic setting, we performed three tests using different training sets. The results of these tests are outlined below.

5.1.1 Fully representative training data

First, we simulated a representative data set by training the classifier on samples from all the light curves in our data set. GP regression was performed on the original 82 light curves. From these, 2000 simulated light curves were generated for each of the 11 classes, to create a balanced training set. These simulated light curves were 8 h long with 100 evenly spaced data points. Wavelet feature extraction was then performed on each of the simulated light curves resulting in 400 wavelet coefficients. After performing PCA on these coefficients, the 100 most important components were selected and used as features. The data set containing 2000 objects per class was then randomly split into 75 per cent training data and 25 per cent test data. It is important to note here that 2000 light curves were simulated for all classes. This includes the classes with only one object (e.g. KN) enabling us to do this train-test split across all eleven classes. While this will lead to overfitting to these single light curve classes, it is still important to keep them in the data set to see which of the other classes the classifier confuses them with.

In order to show the effect of adding contextual information, two separate classifiers were trained. One classifier was trained without any contextual feature (i.e. whether the source was in or out of the Galactic plane) and the other with contextual information as described in Section 4.5. During training, we optimized the following three hyperparameters of the random forests algorithm: the number of trees, the splitting criteria, and the maximum number of features. A grid search was done for the best hyperparameter values using three-fold cross-validation. The best-fitting hyperparameters in this case were found to be 83 trees and a split criterion of entropy with a maximum number of features of 58.

The results are shown in Fig. 3. It can be seen that without any contextual information the classifier confuses the classes of XRB, SNe, Novae, and GRBs. However, after the contextual feature is added, the accuracy for the class of XRBs increases by 16 per cent, SNe increases by 8 per cent, Novae increases by 21 per cent, and GRBs by 8 per cent. Thus, the overall average increase in accuracy is \( \approx 7 \) per cent. It can also been seen that while the overall accuracy increases, some classes such as GRBs and SNe get more confused. This is because these classes were previously confused with novae.

With the introduction of the external information feature, they clearly separate away from novae but are now more likely to be confused with each other instead. However, the overall confusion between classes is reduced. We expect the confusion matrix with contextual information to characterize the performance of the classifier in practice; thus this contextual feature was used in the rest of this work.

It can also be seen that classes with very few original light curves (e.g. KN, RSCVn) have perfect accuracies. This is due to overfitting. The training and test sets for these classes were simulated from a small number of original light curves (between one and three originals). Therefore, these training and test sets are unlikely to have much variation hence making it easy for the classifier to overfit to them.

5.1.2 Non-representative training set

Because our original data set is limited, we expect the results outlined in Section 5.1.1 to represent an idealized case. In practice, we anticipate that any test set would consist of some objects not present in our training set. To construct a more realistic test of the classifier performance with the data currently available, instead of training the classifier on samples from all the original light curves, we trained it on a subset of light curves. As can be seen from Fig. 2, only five classes have greater than four light curves. In order to ensure the training subset contained all classes, we removed light curves from only these five classes. We still included the classes with only one object in both the training and test sets, because these objects can cause confusion between classes that would be artificially removed if they were excluded.

We removed 25 per cent of the light curves in these five classes at random, GP regression was performed on the remaining 75 per cent of the original light curves. From these, 1500 simulated light curves were generated for each of the 11 classes. As before, these simulated light curves were 8 h long with 100 evenly spaced data points. Wavelet feature extraction was then performed on each of the simulated light curves followed by PCA, resulting in 100 features as before. This was used as the training set. A further 500 simulated light curves were drawn from the remaining 25 per cent of the light curves. GP regression, wavelet extraction, and the PCA components from the training set were used to project the high-dimensional feature set down to 100 features as before. This was used as the test set.

\textsuperscript{15}www.simbad.u-strasbg.fr

\textsuperscript{16}https://ned.ipac.caltech.edu

MNRAS 502, 206–224 (2021)
Figure 3. The normalized confusion matrix for wavelet features extracted from 8 h of data. The y-axis shows the true label of the object (true class). The x-axis shows the label that the algorithm predicts for the object (predicted class). The right-hand panel has two colour schemes. The first corresponds to the diagonals. If the values along the diagonal increase they will show in green; if they decrease they will show in red. The second corresponds to the off-diagonals. If the values along the off-diagonals increase they will show in red; if they decrease they will show in green. From the left-hand panel, we see that without any contextual information, the classifier confuses the classes of XRB, SNe, Novae, and GRBs. The classifier is greatly improved with the added feature of the object’s position on the sky. The four classes are no longer confused as most of the off-diagonals are green. The accuracy for the class of XRBs increased by 16 per cent, SNe increased by 8 per cent, Novae increased by 21 per cent, and GRBs by 8 per cent (right-hand panel).

The random forest was trained and the hyperparameters optimized as described in Section 5.1.1.

This was repeated 20 times, each time removing 25 per cent of the light curves at random. The results of this is shown in Figs 4 and 5. It can be seen that the performance of our classifier has decreased by \( \approx 19 \) per cent, which is to be expected given the non-representativeness of the training set. In particular, well-represented classes like AGNs still perform well but classes such as SNe, where the training set is highly diverse with many dissimilar objects, are poorly classified. The overall accuracy averaged over the 20 runs was found to be 78 per cent. We expect the performance of the classifier on real world data, which has similar properties to our data set, to be similar to the performance on these held out data sets.

5.1.3 Single curve testing

To further demonstrate the effect of a non-representative training set, we tested the classifier’s performance on light curves not represented in a training set. We focused on the five classes (AGNs, SNe, XRBs, Novae, and GRBs) that contain four or more light curves.

We trained the classifier on samples from all of the original light curves, except for one. GP regression was thus performed on 82 of the original light curves and from these simulated light curves were drawn, that were 8 h long with 100 evenly spaced data points. Wavelet feature extraction was then performed on each of the simulated light curves followed by PCA. Unlike in the previous two tests, we set the number of features to 20 (from the PCA, which corresponds to retaining 99 per cent of the variability in the data set). One could also perform a hyperparameter optimization over the max feature parameter, as we did in the previous two sections, to find the best number of features to use. While this parameter should ideally be optimized, we did not have the computational resources to optimize

Figure 4. Summary of results from non-representative training set are shown in these box and whisker plots. The bar in each box shows the accuracy for each class averaged over the 20 runs (on each run a classifier was trained on 75 per cent of the total data set and tested on the remaining 25 per cent, this process was repeated 20 times, each time randomizing the training and testing sets). The coloured boxes show the interquartile range. The grey bars (whiskers) show the minimum and maximum accuracies in the 20 runs. It can be seen that the classifier performs poorly for SNe as it has the lowest average accuracy of \( \approx 40 \) per cent.
From Table A1 and Fig. 6, we can see that all the classes except for AGNs have a large variance, this means that the classifier has not adequately learnt the underlying properties of these classes. This is due to the fact that these classes do not have enough examples. It is important to note however, that even with a few examples the classifier can perform well, for example GRBs perform better and have a lower variance than SNe despite SNe having more examples. This is because the four GRB examples in this data set are relatively similar to each other while the 13 SNe examples are not hence the classifier can learn the underlying properties of the GRBs and not the SNe. Therefore, we can expect the classifier to perform well on new objects if they are similar to those in the training data. However, given how small this data set is, it is unlikely that new objects found would be similar thus we do not expect this classifier to generalize well until there is a sufficiently large training set.

As described in Section 2.2.3, the UMAP plots show the distribution of features in high order feature space in two dimensions. It can be seen from Fig. 7 that features from classes where the classifier performs well are relatively separated in feature space. However, data for classes which the classifier confuses overlap in feature space hence the classifier is unable to tell the difference between these objects.

### 5.2 Adding optical data

Simultaneous optical observations do not exist for all the classes in our data set. To show how the addition of multiwavelength data could help in classification, an optical flux measurement had to be simulated. Optical data for four classes, namely: AGNs, XRBs, SNe, and GRBs, were collected from Stewart et al. (2018). The data consisted of single flux measurements in both optical and radio wavelengths for each source. The data set had total of 11,882 measurements of which 11,782 were AGNs. Fig. 8 shows the radio–optical flux distributions for each of the four classes. A two-dimensional Gaussian was used to model the distribution all of the classes except for AGNs which can be clearly seen to be highly non-Gaussian. The Gaussian fits are shown as contours. These Gaussian distributions were used to sample new optical fluxes for the three classes, new optical fluxes were sampled directly from the distribution for AGNs as there are ≈11,000 data points.

Using the class of GRB as an example, the process for simulating simultaneous optical and radio observations is follows. First, a GRB radio light curve was simulated as described previously; the peak flux of this light curve was found. The radio–optical flux distribution was marginalized over, given the peak radio flux, as shown in black in Fig. 8. An optical flux was then drawn from this marginal distribution. Finally, this optical flux was added as an extra feature in the machine learning process. For the class of AGNs, a peak radio flux was found as before, points in the radio–optical flux distribution was then binned, centred on the peak radio flux with a bin width of 0.1 mJy. These binned points, now marginalized over the peak radio flux follow a Gaussian distribution similar to that of the other classes. An optical flux was then drawn from this marginal distribution and added an extra feature in the machine learning process.

The method outlined in Section 5.1.2 was repeated using the four classes mentioned above. The average confusion matrices with and without an optical feature are shown in Fig. 9. It can be seen that the addition of this optical feature improves the performance of the classifier. The average accuracy increases by 5 per cent with the biggest gain seen in the worst performing class, the GRBs, where the number of GRBs misclassified as SNe has dropped by 19 per cent. We note that because our data set of GRBs is drawn from just four
Figure 6. The single curve testing results for the five main classes are summarized in these violin plots. Each violin plot represents a different class as shown. The coloured outline shows the smoothed kernel count distribution of the accuracies. The thick central black line represents the interquartile range. The thin central black line shows the 95 per cent interval. From this, it can be seen that the classifier performs well with AGNs as most of the accuracies are above 80 per cent. It can also be seen that the classifier performs poorly with SNe as most of the accuracies are below 50 per cent. There is notable spread in all the classes, which is a symptom of the small training set and will improve with additional data.

Figure 7. UMAP plots for the four classes of radio transients are shown. UMAP plots show the high-dimensional feature space embedded in two dimensions, hence the $x$ and $y$ axes are arbitrary. The left-hand panel shows the UMAP plot for two classes for which the classifier performs well. The blue points represent feature vectors of AGN objects, and orange points represent the feature vectors of XRB objects. The right-hand panel shows the UMAP plot for two classes for which the classifier performs poorly. The blue points represent feature vectors of SN objects, and orange points represent the feature vectors of GRB objects. It can be seen that features from classes that the classifier does well with are relatively separated in feature space (left-hand panel) whereas features for classes which the classifier confuses overlap in feature space (right-hand panel).

Light curves, this may exaggerate the improvement in performance for this particular class. Thus, more data would be needed to confirm the magnitude of this improvement. Gains in other classes are more modest, in the 2 per cent range. It is interesting to note that with the added optical data, SNe are now slightly (2 per cent) more likely to be confused as GRBs. Thus there is an asymmetry in the off-diagonal terms of the confusion matrix for these two classes. This could be due to the inherent variability of the two classes in our data set, which has a stronger than usual impact due to the small number of light curves in our original (unaugmented) data set. Because there are more SNe
Classifying multiwavelength transients with ML

Figure 8. Optical–radio flux distributions for the classes of AGNs, XRBs, GRBs, and SNe. Two-dimensional Gaussian fits to the distributions of three of the classes are shown with contours. These fit the data for the three classes reasonably well, however it can be seen that the distribution for AGNs is highly non-Gaussian. An example peak GRB radio flux measurement is shown as a black vertical line. The GRB radio–optical flux distribution marginalized over this peak flux is shown in dark grey. Optical fluxes were sampled from these marginals.

in the original data than GRBs, the variability of that class is better covered (and thus likely to be somewhat more realistic) meaning it is more difficult to improve classifier performance with the addition of a simple optical feature. Overall however, these results indicate that including optical data is a promising way to improve classification performance to some degree across all classes. This improvement in performance demonstrates the value of simultaneous optical observations, such as from MeerLICHT in classifying radio transients.

6 CONCLUSIONS

We have presented a general approach for multiwavelength transient classification with machine learning. We have outlined multiple ways to include data from different telescopes, as well as additional information such as source location.

Extending Lochner et al. (2016), we developed a machine learning pipeline for classifying radio transients. We have illustrated how to include contextual information (such as whether or not the source is in the Galactic plane) and simultaneous observations from an optical telescope. This is, to our knowledge, the first case of applying a machine learning formalism for radio transient classification with real data. While our sample size is unfortunately only 82 light curves, we have rigorously investigated the effect of this and have showcased a new data augmentation technique (based on Cui et al. 2016; Revsbech et al. 2018), specifically to address, in a statistically rigorous way, the otherwise perennial issue of a small sample size.

We tested our pipeline using existing radio transient light curves gathered from literature. Because these light curves were few in number, we artificially augmented the data set using GPs. Using GPs as an interpolation and augmentation technique allowed us to split up long light curves into smaller 8-h examples for training, balance the classes, and introduce a small amount of additional variation statistically consistent with the original data (see Appendix D for more details). Features were extracted from this augmented data set using a wavelet decomposition, followed by dimensionality reduction. We found that the wavelet features have much higher performance than a simpler set of features based on the change in flux of the light curve over different time periods (see Appendix B for further details).

We have shown that the classification accuracy increases for some classes when contextual information is included as an additional feature. The improvement is particularly noticeable for novae, objects typically found within the galactic plane, which are otherwise confused with GRBs and SNe.

If the training data are representative of the test data, the performance of the algorithm is excellent, achieving an average accuracy of 97 per cent. However, because our training set is so small, it is highly likely that sources observed by new telescopes such as MeerKAT would not be similar to anything in the training set.

If we remove 25 per cent of the objects from the training set and then test on those removed light curves, we achieve an average accuracy of only 78 per cent. We expect this to be the more realistic performance of the classifier while training sets are small. From the investigation of the effect of dropping out single individual light curves, we find that, while most of the light curves are still classified well, several light curves in the data set are poorly classified. This is because they are dissimilar to anything in the training set. This effect is most pronounced for SNe, which are generally easily confused with other classes such as GRBs or novae. This implies that obtaining training data for such difficult to classify objects should be prioritized as part of modern radio transient surveys.
Finally, we illustrate the effect of including optical data. Unfortunately, simultaneous optical and radio transient data are scarce, so we simulate optical fluxes for the radio light curves, drawing on existing distributions. Using the more realistic ‘dropped out’ training set, we find that the addition of the optical feature dramatically reduces the confusion between SNe and GRBs.

Our extensive analysis shows that while the technique performs well and small training sets can be utilized, it is still limited by the availability of a representative training set. This is difficult to obtain due to few examples of radio transients as well as biases in selection effects, telescope sensitivity, etc. In general, it can be difficult to know when one has enough training examples of a given class to provide reliable classification. Classes exhibiting more intra-class variability generally require more training examples. Our approach outlined in Sections 5.1.2 and 5.1.3 can assist in determining this since the classifier performance will not vary greatly as subsets of data are removed once there are sufficient training examples for that class. Additionally, one should monitor the performance of the classifier as object classes are spectroscopically confirmed. While small training sets are a challenge in any machine learning setting, we do expect that the algorithm will scale well and achieve good performance with real data, with classification performance and reliability increasing as more training data are added. For use on a modern telescope such as MeerKAT, we recommend the classifier be used in online mode where new training examples are added continuously, thus improving performance of the algorithm.

These results indicate that by including multiwavelength information and making use of a sophisticated machine learning approach, we can expect accurate classification of radio transients with even a small training set created with early MeerKAT and MeerLICHT data.

ACKNOWLEDGEMENTS

We acknowledge support from SKA Africa and the National Research Foundation (NRF) towards this research. Opinions expressed and conclusions arrived at, are those of the authors and are not necessarily to be attributed to the NRF. We acknowledge support from STFC for UK participation in LSST through grant ST/N00258X/1 and travel support provided by STFC for UK participation in LSST through grants ST/L00660X/1 and ST/M00015X/1. HVP was supported by the European Research Council under the European Community’s Seventh Framework Programme (FP7/2007-2013)/ERC grant agreement no. 306478-CosmicDawn. PAW acknowledges support from the University of Cape Town. OL acknowledge STFC grant ST/R000476/1.

DATA AVAILABILITY

The data underlying this article are available in the Zenodo Digital Repository, at https://dx.doi.org/10.5281/zenodo.4035188.
Classifying multiwavelength transients with ML

Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Perez L., Wang J., 2017, preprint (arXiv:1712.04621)
Pietka M., Fender R. P., Keane E. F., 2015, MNRAS, 446, 3687
Quinlan J. R., 1993, C4.5: Programs for Machine Learning, Morgan Kaufmann Publishers Inc., San Francisco, CA
Rasmussen C. E., Williams C. K. I., 2005, Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press, Cambridge
Rebbapragada U., Lo K., Wagstaff K. L., Reed C., Murphy T., Thompson D. R., 2012, in Griffin E., Hanisch R., Seaman R., eds, New Horizons in Time Domain Astronomy. Cambridge University Press, Cambridge, p. 397
Revsbech E. A., Trotta R., van Dyk D. A., 2018, MNRAS, 473, 3969
Richards J. W. et al., 2011, ApJ, 733, 10
Romano R., Aragon C., Ding C., 2006, ApJ, 665, 1246
Stewart A. J., Muñoz-Darias T., Fender R. P., Pietka M., 2018, MNRAS, 479, 2481
van der Maaten L., Hinton G., 2008, J. Mach. Learn. Res., 9, 2579
Vargas R., Mosavi A., Ruiz R., 2017, Advances in Intelligent Systems and Computing, 5, 1

APPENDIX A: TABLES

Table A1 shows the results of the individual dropout test.

| Name       | Acc |
|------------|-----|
| NGC 7213   | 0.0 |
| 0850−121   | 0.0 |
| 0954+658   | 0.206 |
| 0224+671   | 0.784 |
| 2005+403   | 0.824 |
| NRAO530    | 0.931 |
| 2223−052   | 0.951 |
| 1413−135   | 0.961 |
| 0528−134p  | 0.98 |
| 1622−297   | 1.0 |
| CTA102     | 1.0 |
| 0336−019   | 1.0 |
| 3C 345     | 1.0 |
| B0605−085  | 1.0 |
| 3C 454.3   | 1.0 |
| 1328+254   | 1.0 |
| 3C 120     | 1.0 |
| 3C 273     | 1.0 |
| 0458−020   | 1.0 |
| 3C 279     | 1.0 |
| 1237+049   | 1.0 |
| 0851+202   | 1.0 |
| 2200+420   | 1.0 |
| 0528+134   | 1.0 |
| PKS2004−447| 1.0 |
| 1803+784   | 1.0 |
| 0954+65    | 1.0 |
| 1749−096   | 1.0 |
APPENDIX B: FLUX FEATURES

B1 Flux feature extraction

This section details an investigation of a set of features that were ultimately not used in this paper. We include it here because the features, a simple difference in flux over different time steps, are likely to be an obvious choice for this problem and we thus consider it useful to demonstrate that they are likely inadequate for transient classification.

The feature vector, $\Delta F$, was defined to be

$$\Delta F = F_t - F_{t_0}, \quad (B1)$$

where $\Delta F$ is the difference in flux between a reference flux, $F_{t_0}$, chosen at random from the light curve and the flux, $F_t$, at time $t > t_0$.

In anticipation of a fast imager on MeerKAT, we chose the minimum difference between successive flux measurements to be $2$ s. The maximum time difference was chosen to be $3$ months to account for the objects that vary on very long time-scales such as AGNs. The complete set of $t$ where chosen to be

$$t = [2 \text{ s, } 1 \text{ min, } 5 \text{ min, } 10 \text{ min, } 30 \text{ min, } 1 \text{ h, } 2 \text{ h, } 4 \text{ h, } 6 \text{ h, } 8 \text{ h}, \ 12 \text{ h, } 18 \text{ h, } 1 \text{ d, } 2 \text{ d, } 4 \text{ d, } 1 \text{ d, } 2 \text{ week, } \ 3 \text{ week, } 1 \text{ month, } 1.5 \text{ month, } 2 \text{ month, } 3 \text{ month}]. \quad (B2)$$

The feature extraction method was then as follows. First, GP regression was performed on the original data set. A reference time, $t_0$, was then drawn at random to be somewhere within the curve. The GP was then sampled at this $t_0$ to obtain an $F_{t_0}$. The GP was then sampled at points $(t_0 + t)$. $F_t$ was then subtracted from these fluxes to obtain the feature vector $\Delta F$. This process was repeated multiple times for each light curve, each time generating a random $t_0$. This was done to simulate the fact that the transient may be detected at any point on the light curve.

B2 Results

Once the flux features were extracted as described in Section B1, different subsets of the total feature set was used to train different classifiers. The subsets used were created by truncating the features at different time-scales, i.e. a classifier was trained on all features up to a maximum of $5$ min, then another was trained on all features up to a maximum of $10$ min and so on for all $t$ in equation (B2). The accuracy of each of these classifiers is shown in Fig. B1.

It can see from Fig. B1 that the classifier performs well on long time-scales. We would like to investigate is how well the classifier performs on short time-scales for each of this classes. It can be seen from Fig. 2 that the number of FS light curves in the data set goes to zero, hence $8$ h is the longest time-scale at which we have a complete set of classes. Thus, the time vector from Section B1 is changed to

$$t = [2 \text{ s, } 1 \text{ min, } 5 \text{ min, } 10 \text{ min, } 30 \text{ min, } 1 \text{ h, } 2 \text{ h, } 4 \text{ h, } 6 \text{ h, } 8 \text{ h}].$$

Fig. B2 shows the normalised confusion matrixes for flux features extracted from $8$ hours of data.
Classifying multiwavelength transients with ML

Figure B2. The normalized confusion matrix for flux features extracted from 8 h of data. The y-axis shows the true label of the object (true class). The x-axis shows the label which the algorithm predicts for the object (predicted class). It can be seen that while contextual information does improve the performance of the classifier, this method of feature extraction does not do well to separate the classes.

APPENDIX C: VARIABLE AND TRANSIENT CONFUSION MATRICES

From Fig. 1, it can be seen that radio transients can be split into two types: transient that have multiple outbursts on short time-scales (e.g. FS) and transients that have one outburst that then decays over longer time-scales (e.g. SNe). These two groups are called variables and transients, respectively. The data set was split into these two groups. A classifier was trained on the binary classification of transient versus variable. Classifiers were then trained on the individual groups of variables and transients. The results of this is shown in Figs C1 and C2.

It can be seen from Fig. C1 that the classifier can distinguish between variables and transients extremely well, with an overall accuracy of 98.5 per cent. From Fig. C2, we can see that by making this split we increase the accuracy of the XRB and SN classifications slightly when compared to Fig. 3. The class of GRB, however, sees no improvement. As with Fig. 3, SNe are still being confused with TDEs. It can be seen that this split is unnecessary as it achieves little to no improvement in the performance of the classifier.

Figure C1. The normalized confusion matrix for wavelet features extracted from 8 h of data. The y-axis shows the true label of the object (true class). The x-axis shows the label which the algorithm predicts for the object (predicted class). It can be seen that the classifier can distinguish between variables and transients extremely well.
Figure C2. The normalized confusion matrix for wavelet features extracted from 8 h of data. The y-axis shows the true label of the object (true class). The x-axis shows the label which the algorithm predicts for the object (predicted class). By making this split, we increase the accuracy of the XRB and SN classifications. The class of GRB, however, sees no improvement. As with the previous methods, SNe are still being confused with TDEs. It can be seen that this split is unnecessary as it achieves little to no improvement in the performance of the classifier.

APPENDIX D: INVESTIGATION ON AUGMENTATION

In this appendix, we investigate the value of the different components of our augmentation procedure based on GPs. We first redo our non-representative analysis, where 25 per cent of the light curves from each class are entirely excluded from the training set. This time however, we only draw one 8 h sample from each light curve, instead of many. We then explore the impact of adding more examples from the long light curves, balancing the data, and finally sampling from the GP to fully understand the effect of augmentation. We note that because we have few light curves, the variance in each of our tests (which we repeat 100 times) is large. Thus, we find that while data augmentation essentially always improves performance, the effect it has dramatically depends on the exact training data used.

We restricted our data set to the five main classes used in Section 5. Thus, our data set contained 30 AGNs, 14 XRBs, 13 SNe, 8 Novae, and 4 GRBs. GP regression was performed on these and one 8 h observation was sampled from each of the light curves resulting in an unbalanced data set containing a total of 69 light curves. A classifier was then trained on this data set. The results are shown in Fig. D1.

Fig. D1 shows degraded performance in comparison to the original Fig. 5. Because we leave out the additional classes in this case, we could actually expect performance to be even worse. It does show however that AGNs are still very straightforward to classify owing to their lack of variation over the time-scale we are interested in. We argue that augmentation can improve this base performance, particularly for classes with few examples. We test this by performing three experiments using different methods to increase the number of light curves in our data set.

The first experiment investigates the effect of splitting the light curves up into many 8 h ‘chunks’ to increase the number of light curves. In the second experiment, we increase the number of light curves while also balancing the number of objects per class. Finally, we investigate the effect of sampling from the GPs to add additional curves.
Figure D2. The increase in accuracy per class and the overall increase in accuracy as a function of data set size relative to the original data set. We can see that initially the performance of the classifier improves on all classes except for AGNs. AGNs are the majority class thus the classifier learns that it is beneficial to label more objects as AGNs. As we increase the number of objects in the smaller classes, the classifier learns to better classify them leading to less objects being mis-classified as AGNs.

D1 The effect of splitting light curves

The radio light curves from the five main classes mentioned above are all relatively long; therefore, we could increase the training set size by splitting the light curves into 8 h pieces. The GPs were used as a statistically rigorous method for interpolation rather than a method for augmentation, because we sample these ‘chunks’ from the mean function. We extract 8 h chunks in sets of 1, 2, 5, 10, 30, and 60 per light curve, effectively increasing the training data each time. It is important to note here that the original ratio of the class numbers is the same, i.e. each data set is still imbalanced with the total number of light curves increased by $2 \times$, $5 \times$, etc.

A classifier was trained on the each of the sampled data sets. This was then repeated 100 times each time randomizing the training and test sets as described in Section 5.1.2. It is important to note that the test set was balanced, consisting of around 200 samples per class drawn from the test light curves. It is only the training data that is imbalanced. The percentage increase in accuracy was then calculated for each of the data sets with respect to the data set containing one sampled light curve per original light curve. This easily shows whether or not there is an increase in performance independent of the intrinsic accuracy between runs. The average percentage increase in accuracy is shown in Fig. D2. From this figure, it can be seen that the accuracy of all classes except for AGNs increase. AGNs are the class with the highest number of objects, thus the algorithm learns that it is beneficial to label more objects as AGNs; hence AGNs have a high accuracy. As we increase the number of objects in the smaller classes the classifier learns to better classify them, less objects are mis-classified as AGNs and hence the accuracy of AGNs decrease. Fig. D2 also shows the general trend that increasing the number of light curves in the data set leads to an increase in overall accuracy.

D2 The effect of increasing the number of light curves while balancing the classes using GPs

The experiment proceeded as follows: GP regression was performed on each of the original light curves as before. We then sampled 8 h chunks from the means of these GPs this time creating balanced data sets, where each class contained 30, 100, 500, and 1000 light curves, respectively. As before, a classifier was trained 100 times on each of the sampled data sets, randomizing the training and test sets as described in Section 5.1.2. The percentage increase in accuracy was then calculated for each of the data sets with respect to the data set containing one sampled light curve per original light curve. Fig. D3 shows the overall increase in accuracy over the 100 runs plotted in blue.

Finally, we explore the impact of adding variation by drawing samples from the GP instead of just using the mean. These are
plotted in grey in Fig. D3. From these we can see that both splitting up the light curves using the means of the GPs as well as sampling from the entire GP improves the performance of the classifier. Using random samples drawn from the GP instead of just the mean improves the accuracy by \( \approx 2 \) per cent on average and slightly reduces the variance in the accuracy between runs. Comparing this to the previous experiment we see that for similar sized data sets, the balanced data set outperforms the unbalanced data set by 3–5 per cent. For this particular problem, augmentation increase accuracy by around 11 per cent on average and could improve performance by considerably more depending on the exact training set used. Thus, we recommend the use of augmentation when classifying small, diverse populations of transients.

**APPENDIX E: t-SNE – AN ALTERNATIVE VISUALIZATION TOOL**

We ran t-SNE for the data in Fig. 7, a Euclidean metric was used when estimating the manifold. We found that while the plots are structurally different, they demonstrate the same conclusion. Fig. E1 shows the t-SNE versions of the plots shown in Fig. 7.

![t-SNE plots](example.png)

**Figure E1.** An alternative to Fig. 7, t-SNE plots for the four classes of radio transients are shown. t-SNE plots show the high-dimensional feature space embedded in two dimensions. As in Fig. 7, the plots show good separation between AGNs and XRBs but much more overlap between SNe and GRBs, which are more difficult to classify.

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