AUTOMATIC CLASSIFICATION OF TIME-VARIABLE X-RAY SOURCES

Kitty K. Lo\textsuperscript{1,2}, Sean Farrell\textsuperscript{1,2}, Tara Murphy\textsuperscript{1,2,3}, and B. M. Gaensler\textsuperscript{1,2}

\textsuperscript{1} Sydney Institute for Astronomy, School of Physics, The University of Sydney, Sydney, NSW 2006, Australia
\textsuperscript{2} ARC Centre of Excellence for All-Sky Astrophysics (CAASTRO), Redfern, NSW 2016, Australia
\textsuperscript{3} School of Information Technologies, The University of Sydney, Sydney, NSW 2006, Australia

Received 2013 March 17; accepted 2014 February 23; published 2014 April 10

ABSTRACT

To maximize the discovery potential of future synoptic surveys, especially in the field of transient science, it will be necessary to use automatic classification to identify some of the astronomical sources. The data mining technique of supervised classification is suitable for this problem. Here, we present a supervised learning method to automatically classify variable X-ray sources in the Second XMM-Newton Serendipitous Source Catalog (2XMMi-DR2). Random Forest is our classifier of choice since it is one of the most accurate learning algorithms available. Our training set consists of 873 variable sources and their features are derived from time series, spectra, and other multi-wavelength contextual information. The 10 fold cross validation accuracy of the training data is $\sim$97% on a 7 class data set. We applied the trained classification model to 411 unknown variable 2XMM sources to produce a probabilistically classified catalog. Using the classification margin and the Random Forest derived outlier measure, we identified 12 anomalous sources, of which 2XMM J180658.7$-$500250 appears to be the most unusual source in the sample. Its X-ray spectra is suggestive of a ultraluminous X-ray source but its variability makes it highly unusual. Machine-learned classification and anomaly detection will facilitate scientific discoveries in the era of all-sky surveys.

Key words: catalogs – methods: statistical – X-rays: general

Online-only material: color figures, machine-readable table

1. INTRODUCTION

The identification of variable and transient astrophysical sources will be a major challenge in the near future across all wavelengths. The advent of facilities such as the Large Synoptic Survey Telescope (LSST) in optical (Tyson 2002), the Square Kilometre Array (SKA) in radio (Cordes et al. 2004) and the extended ROentgen Survey with an Imaging Telescope Array (eROSITA) in X-rays (Merloni et al. 2012), will enable the next generation of all-sky time-domain surveys. Many types of transients and variable sources are currently known, such as supernovae, cataclysmic variables (CVs), X-ray binaries (XRBs), flare stars, gamma-ray bursts (GRBs), tidal disruption flares, and future time-domain surveys will likely uncover novel source types. The large number of sources to be surveyed makes identifying interesting transients a challenging task, especially since timely multi-wavelength follow-ups will be critical for fulfilling the transient science goals. To this end, we envision that automatic classification will be a crucial part of the processing pipeline (Murphy et al. 2013).

Here, we demonstrate the feasibility of using time series and contextual information to automatically classify variable and transient sources. We used data from the X-ray Multi-Mirror Mission-Newton (XMM-Newton) because there has not been previous studies on this data set using automatic classification algorithms and because the time series for many of the sources are readily available, thereby enabling us to investigate the efficacy of a classifier built using solely time-domain information. Automatic classification is a similar problem across all wave-lengths and we expect that the techniques used in this paper can be readily adapted for data sets in other wave-bands.

The Second XMM-Newton Serendipitous Source Catalog Data Release 2 (2XMMi-DR2) was the largest catalog of X-ray sources (Watson et al. 2009) at the time it was released, but has since been surpassed by 2XMMi-DR3 and 3XMM. In this study, we used 2XMMi-DR2 and kept DR3 as a verification sample. There have been previous attempts to classify the unidentified sources in 2XMMi (Pineau et al. 2011; Lin et al. 2012). The traditional method is to cross-match the unknown sources with catalogs in other wavelengths (e.g., Sloan Digital Sky Survey, Two Micron All Sky Survey (2MASS)) and then use expert knowledge to draw up classification rules. For example, one powerful discriminant is the ratio of the optical to X-ray flux for separating active galactic nuclei (AGN) and stars. In the scheme used by Lin et al. (2012), sources whose positions coincide with the centers of galaxies are deemed to be AGN. Manually selected classification rules often have their basis in science and are usually comprehensible to other experts. This method works well when there are only a few pieces of information to be processed (e.g., optical to X-ray flux), but becomes intractable when there are many disparate sets of information. In machine learning, each piece of information is translated into either a real number or a categorical label known as a feature. Machine learned classification excels at finding subtle patterns in data sets with a large number of features.

Machine learned classification has been used extensively in astronomy. In X-ray astronomy, McGlynn et al. (2004) used oblique decision trees to produce a catalog of probabilistically classified X-ray sources from ROSAT. Since that study, there have been many advances in automatic classification techniques. Ensemble algorithms such as Random Forest (RF) have replaced single decision trees as the state-of-the-art. RF has been successfully used in astronomy for the automatic classification of variable stars (Richards et al. 2011; Dubath et al. 2011) and the photometric classification of supernovae (Carliles et al. 2010). In optical astronomy, there are efforts to incorporate automatic classification in the processing pipelines of current and planned surveys (Saglia et al. 2012; Bloom et al. 2012; Djorgovski et al. 2012).

Feature representation is an important issue in light curve classification. Since light curves are rarely observed with exactly the same cadences, they need to be transformed into structured
feature sets before different sources can be compared. Various light curve feature representations have been used in astronomy. For example, Matijević et al. (2012) transformed the light curves of each Kepler eclipsing binary into a set of 1000 observations by fitting and then interpolating the observations. However, this method only works for a very homogenous set of light curves. Other studies use a restrictive set of variability measures. In Hofmann et al. (2013), X-ray sources in M31 are placed into two light curve classes—highly variable or outbursts. This method has limited descriptive power for the variety of time-variability behaviors. In contrast, Rimoldini et al. (2012) extracted a large number of features from each light curve in the Hipparcos catalog and used RF and Bayesian networks to automatically classify ~6000 unsolved optical variables. They achieved a misclassification rate of less than 12% and this is the methodology for feature representation that we have used.

In this paper, we present the results of using the RF algorithm to classify variable sources in 2XMMi-DR2. In Section 2, we describe the 2XMMi-DR2 data set and the data processing we performed. In Section 3, we describe the RF algorithm. In Section 4 we present the classification results using only time-series features and in Section 5, we show how the classification accuracy increases with the inclusion of contextual features. Our main result, a table of probabilistically classified 2XMMi variable sources, is presented in Section 6. In Section 7 we present a method for selecting anomalous sources and briefly describe one of the interesting anomalous source. Finally, in Section 8 we discuss the limitations and future prospects of machine learned classification.

2. THE 2XMM VARIABLE SOURCES

The 2XMMi-DR2 catalog consists of observations made with the XMM-Newton satellite between 2000 and 2008 and covers a sky area of about 420 deg². The observations were made using the European Photon Imaging Camera (EPIC) that consists of three CCD cameras—pn, MOS1, and MOS2—and covers the energy range from 0.2 keV to 12 keV. There are 221,012 unique sources in 2XMM-DR2, of which 2267 were flagged as variable by the XMM processing pipeline (Watson et al. 2009). The variability test used by the pipeline is a χ² test against the null hypothesis that the source flux is constant, with the probability threshold set at 10⁻⁵.

2.1. Data Processing

In this paper, a detection refers to a light curve in an epoch made by one camera. Each detection in our sample has an associated light curve which consists of background subtracted count rates, count rate errors, background count rates, background errors, and time stamps. A source can be detected in multiple epochs, and in each epoch there are typically three detections, one by each of the pn, MOS1 and MOS2 cameras. The exposure time per detection ranges from a few kiloseconds to over 100 ks (Figure 1). The bin widths are in multiples of 10 s and are large enough such that there are a minimum of 18 counts bin⁻¹ and 5 counts bin⁻¹ for the pn and MOS detectors, respectively. To ensure that all the variability in the light curve comes from the source and is not due to background flares or instrumental errors, we filtered out points likely to contain errors. First, we removed all points that lie outside the good time intervals (GTIs). GTIs are time periods where monitored parameters, such as spacecraft attitude stability and background particle levels, are within acceptable levels. Second, we removed all points where the fraction of time exposed, $F_{\text{exp}}$, was <0.9. Count rates determined during a low $F_{\text{exp}}$ measurement are not reliable. Third, we removed points with zero error rates. Since an error of zero is not realistic, it indicates some error in the data processing or the observation. After the filtering step, we removed sources from the sample with less than 15 data points in the light curve. Table 1 is a breakdown of the sources in our sample. In total, we excluded 983 sources from further considerations.

2.2. Classified Sample

For our training set we used the classifications for each discrete variable 2XMMi-DR2 source as determined by S. Farrell et al. (in preparation). While the classification methodology will be discussed in detail in S. Farrell et al. (in preparation), we summarize the process as follows. First, the pipeline produced images, spectra, and light curves were manually inspected using the products available on the Leicester Database and Archive Service Web sites. Spurious detections were identified, primarily through examination of the images, and summarily discarded. Detections of extended sources were also discarded (e.g., supernova remnants, galaxy clusters etc.) as any variability detected from these sources within a single XMM-Newton observation would have to be spurious. In this manner, we discarded 924 out of the original 2267 variable sources as spurious.

| Table 1: The Variable 2XMMi Sample |
|-----------------------------------|
| Sources                          |
|-----------------------------------|
| Total excluded from our sample    | 983 |
| Spurious                         | 924 |
| Classified: not enough data points| 14  |
| Classified: classes with few sources| 17  |
| Unidentified: not enough data points| 28  |
| Total in our sample               | 1284|
| Classified: in the training set   | 873 |
| Unidentified: in the test set     | 411 |
| Total variable sources            | 2267|

4 http://www.ledas.ac.uk/
The nature of the remaining 1343 real variable sources was determined by searching for matches around the source positions in the SIMBAD astronomical database and the NASA/IPAC Extragalactic Database (NED), and through a shallow review of the literature. The bulk of these sources (44%) were associated with stars, with the rest associated with the centers of galaxies (i.e., AGN; 7%), XRBs (6%), CVs (6%), ultraluminous X-ray sources (ULXs; 1%), GRBs (1%), and super soft sources (SSS; 1%). A very small number (representing ~1%) were associated with planets (Jupiter and Saturn), extragalactic globular clusters, and magnetars. The remaining sources, comprising 33% of the real variable source sample, did not have a match in either SIMBAD or NED and are thus unidentified. The training set thus contains 873 sources in seven classes: AGN, CVs, GRBs, XRBs, SSSs, stars, ULXs, and XRBs, with the unidentified sources represented by the symbols "\( \sim \)".

---

\(^5\) http://simbad.u-strasbg.fr/simbad/

\(^6\) http://nedwww.ipac.caltech.edu/

---

**Figure 2.** Example lightcurves for the seven types of X-ray sources in our training set.
sources not included. Table 2 shows a breakdown of the number of sources and detections we have in the classified training set and Figure 2 shows examples of light curves from each class.

AGN are the central regions of galaxies believed to contain supermassive black holes. X-ray emission from AGN is mainly due to inverse Compton scattering and typically follows a power-law spectrum (Longair 2011). We included different types of stars under the “star” category, including flare stars, binaries, pre-main sequence stars and young stellar objects. Late-type flare stars produce X-ray emission from magnetic reconnection in their coronae (Benz & Güdel 2010). A CV is a binary system in which a white dwarf accretes from a companion star. The typical orbital periods of CVs are between 75 minutes and 8 hours. CVs can be magnetic (mCV) or non-magnetic; the former are also known as polars or intermediate polars. X-ray emission from non-magnetic CVs is mainly due to low temperature thermal plasma emission from shocks formed when material accretes onto the white dwarf. In mCVs, the accretion disk is suppressed by the magnetic field and the X-ray emission arises from the boundary of the shock of the collimated accretion flow. XRBs are binary systems where the accreting compact object is a black hole or neutron star. The donor star in a high-mass XRB is usually a massive O- or B-type star, or a Be star while the donor star in a low-mass XRB can be a main-sequence star, a white dwarf or a red giant. Both subtypes of XRBs are included in this category. ULXs are objects with X-ray luminosities exceeding that of a stellar mass black hole accreting at the Eddington limit. They are located within galaxies but not in the nucleus regions. SSSs, as their name suggests, are characterized by their extremely soft (peaking at <0.5 keV) spectra. The accepted paradigm for their nature is that of a white dwarf binary with steady nuclear burning (Kahabka & van den Heuvel 2006). Lastly, the GRBs we are referring to here are afterglow emission from long GRBs, which are believed to be the core collapses of massive stars.

3. CLASSIFICATION METHOD

3.1. Introduction

The machine learning technique we use here is known as supervised classification (Duda et al. 2001). Supervised classification uses a set of labeled training examples to construct a prediction model. In Section 2, we described the method used to construct the training set. In this section we will explain the classification algorithm in detail.

3.2. Random Forest

RF is an ensemble supervised classification algorithm developed by Breiman (2001). In the training phase, the algorithm builds an ensemble of decision trees. Each tree is built using a bootstrap sample from the training set, i.e., for S samples in the training set, the algorithm randomly picks S samples with replacement to create the training set for each tree. To construct a decision tree, training samples are split at each node (a node is where classification decisions are made) and this process iterates until all the training samples at the node belong to the same class. The feature used at each node is the one that produces the highest decrease in Gini impurity, as calculated using the equation:

\[ G = \sum_{i=1}^{m} f_k (1 - f_k), \]

where \( m \) is the number of classes and \( f_k \) is the fraction of sources which belong to class \( k \). The Gini impurity becomes zero when all the sources in a node are of the same type. In RF, a small subset of features (typically only a small fraction of the total number of features) are randomly chosen to be considered at each node. To predict the class of a new sample, each decision tree in the ensemble votes for a class and the output class is the one with the most votes.

RF is one of the most accurate classification algorithms available (Caruana & Niculescu-mizil 2006). It can handle large data sets with large number of features. RF can generalize without overfitting; an overfitted model is one that describes noise rather than the true underlying relationship between features. It is also simple to optimize, since there are only two parameters to adjust—the number of trees and the number of variables to use at each node. We used the R package (R Core Team 2013) randomForest (Liaw & Wiener 2002) for the experiments performed in this paper. Using the tuning function tuneRF in the randomForest package, we found that the optimal number of variables to use at each node is 9. To find the optimal number of trees, we repeated the experiment with different number of trees, and found 500 trees was optimal.

3.3. Unbalanced Training Set

Our training set, as summarized in Table 2, is heavily unbalanced. Stars, the most abundant class, outnumber GRB, the rarest class by around 200 to 1. Heavily unbalanced training sets can degrade the performance of a machine learned classification algorithm. To ameliorate this issue, we oversampled the two most under-represented classes—GRB and SSS, using the SMOTE algorithm (Chawla et al. 2002). SMOTE creates synthetic minority class samples by using the k-nearest neighbors and has been shown to be more robust than simply oversampling the minority class with replacement. We used the SMOTE implementation in the DMwR package (Torgo 2010) in R to oversample the GRB class by ten fold and the SSS class by four fold.

3.4. Class Membership Probabilities

Class membership probabilities can be more informative than discrete class labels. The former provides information on the degree of confidence of the classification, and allows the user to set cutoffs for selecting their class of interest based on the desired level of reliability and completeness. RF can provide class membership probabilities in the form of the fraction of votes in the ensemble given for the class. In this paper, we report all results as class membership probabilities.
4. CLASSIFICATION WITH TIME SERIES FEATURES

The variable X-ray source sample allows us to investigate the usefulness of variability information in classification. In this section, we describe the time series features we extracted from the X-ray light curves and report on the accuracy of the RF classifier trained using only time series features. Table 3 is a summary of the general light curve characteristics of each source type. Although we cannot arrive at a definitive classification solely using the light curves, variability information can narrow down the potential classes. For example, a source with periodic variability is highly unlikely to be an AGN, but it could be a CV or an XRB.

Based on the expected variability characteristics, we extracted four types of light curve features—periodic features, likelihood of power law decay, flares and statistical features such as fractional variability. Table 4 is a summary of the time series features. We discuss each feature in detail in the following sections.

### 4.1. Periodic Features

Some CVs and XRBs display periodicities on timescales of minutes to hours, less than the typical length of our observations (Israel et al. 2002; Hearn & Richardson 1977). This suggests the frequency domain can inform our classification. We used the generalized Lomb–Scargle periodogram from Zechmeister & Kürster (2009) to represent the frequency domain information. The advantage of this technique over a conventional Fourier transform is that it can handle unevenly sampled data. For evenly sampled light curves, this would be unnecessary. However, due to the filtering process, our light curves may be missing data points. The generalized periodogram is equivalent to fitting functions of the form $y = a \cos \omega t + b \sin \omega t + c$. The inclusion of the offset $c$ makes it more general than the original Lomb–Scargle periodogram (Lomb 1976). Finding the best fit translates to minimizing the squared difference between the data at time $i$, $y_i$, and the model $y(t_i)$ represented by the $\chi^2$ function:

$$
\chi^2 = \sum \frac{(y_i - y(t_i))^2}{\sigma_i^2},
$$

where $\sigma_i^2$ is the estimated variance at time $i$.

The periodogram can be written as:

$$
P(\omega) = \frac{N - 1}{2} \frac{\chi_0^2 - \chi^2(\omega)}{\chi_0^2},
$$

where $\chi_0$ is the squared deviation of $y_i$ from the mean.

Equation (3) has been normalized by the factor $(N - 1)/2$ ($N$ is the number of measurements in the time series) so that if the data are pure noise, then the expected periodogram value is 1.

### Table 3

| Class       | Light Curve Characteristics                                      |
|-------------|-----------------------------------------------------------------|
| AGN         | Flickering, stochastic aperiodic variability                    |
| CV          | Some sources display periodicity                                 |
| GRB         | Power law decay                                                 |
| SSS         | Typically constant, occasional variability                      |
| Star        | Flares and bursts lasting minutes                                |
| ULX         | Typically constant, occasional variability                      |
| XRB         | Some sources display periodicity, flickers and flares          |

### Table 4

| Feature                  | Description                                                                 |
|--------------------------|-----------------------------------------------------------------------------|
| Lomb-Scargle_amp1        | Amplitude of the best-fitting sine function at the highest Lomb–Scargle periodogram peak |
| Lomb-Scargle_amp2        | Amplitude of the best-fitting sine function at the second highest Lomb–Scargle periodogram peak |
| Lomb-Scargle_period1     | Period in seconds corresponding to the highest Lomb–Scargle periodogram peak |
| Lomb-Scargle_period2     | Period in seconds corresponding to the second highest Lomb–Scargle periodogram peak |
| Lomb-Scargle_FAP1        | False alarm probability of the highest Lomb–Scargle periodogram peak         |
| Lomb-Scargle_FAP2        | False alarm probability of the second highest Lomb–Scargle periodogram peak   |
| Powerlaw_C               | Parameter $C$ in the best-fit power-law model $y(t) = F_0(t - \tau_0)^{-C}$   |
| Powerlaw goodness of fit | Reduced $\chi^2$ statistics of the exponential model                         |
| Flare_nums               | Number of flares found                                                       |
| Flare_amp                | Amplitude of the strongest flare                                             |
| Flare_duration           | Duration in seconds of the strongest flare                                   |
| Amplitude                | $0.5 \times (\text{Max(count)} - \text{Min(count)})$                         |
| Standard_deviation       | Standard deviation of the counts                                             |
| Beyond1Std              | Percentage of observations that lie beyond one standard deviation from the weighted mean |
| Flux ratio mid 20%       | Ratio of flux in the 60th to 40th percentiles over 95th to 5th percentiles   |
| Flux ratio mid 35%       | Ratio of flux in the 67.5th to 32.5th percentiles over 95th to 5th percentiles |
| Flux ratio mid 50%       | Ratio of flux in the 75th to 25th percentiles over 95th to 5th percentiles   |
| Flux ratio mid 65%       | Ratio of flux in the 82.5th to 17.5th percentiles over 95th to 5th percentiles |
| Flux ratio mid 80%       | Ratio of flux in the 90th to 10th percentiles over 95th to 5th percentiles   |
| Skew                     | Skew of the distribution of count rates; calculated using the python function scipy.stats.skew |
| Max slope                | Maximum slope of adjacent observation points (counts s$^{-1}$)               |
| Med_abs_dev              | Median of the absolute value of the deviation from the median               |
| Med_buffer_range_percentage | Percentage of measurements within 20% of the median                        |
| Percentage_amp_diff      | Maximum difference between a measurement and the median as a percentage of the median |
| Percentile_diff          | Count rate at the 98th percentile minus the count rate at the 2nd percentile |
| Modulation Index         | Variance/weighted mean                                                      |
| Fractional var           | Fractional rms variability, calculated as $\sqrt{\frac{1}{N} \sum (y_i - \bar{y})^2 - \frac{1}{N} \sum \sigma_i^2}$, where $y_i$ is the count rate at time $i$, $\sigma_i$ is the error of the count rate at time $i$, $\bar{y}$ is the average count rate and $N$ is the number of observations in the time series |
This equation has an analytical solution (Equation (5) in Zechmeister & Kürster 2009) that we used to calculate the periodogram value. The false alarm probability (FAP) is also included in our feature set as a way to capture the significance of the periodogram value. FAP is calculated using:

$$\text{FAP} = 1 - \left[ 1 - \left( 1 - \frac{2P}{N-1} \right)^{\frac{N-1}{2}} \right]^M,$$

where $M$ is the number of peaks in the periodogram. This relies on an implicit assumption that the noise in the flux is Gaussian.

For our classification experiments, we only used the two highest peaks in the periodogram. For each peak, we extracted the amplitude of the best-fitting sine function ($\sqrt{a^2 + b^2}$), the period in seconds and the FAP. Figure 3 shows a plot of the first two of these three values. To ensure the second peak is truly distinct from the first, we eliminated values immediately adjacent to the highest peak and found the second highest peak from the remaining frequency bins.

### 4.2. Fit to a Power-law Model

The identifying feature of a GRB afterglow light curve is a power-law function. We fitted a power-law function of the form $y(t) = F_0 (t - t_0)^{-c}$ to the light curves and used the parameters $C$ of the best-fitting model as classification features. The fitting procedure also determined $F_0$ and $t_0$ but these were not used as classification features. We used the `curvefit` function from the Python package `scipy` to perform the least squares nonlinear fit. This process assumes the input errors are Gaussian, which is not always satisfied due to low count rates. To circumvent this issue, we binned the data to coarser time bins such that the average number of counts per bin was at least 20. To estimate the goodness of fit, we calculated the $\chi^2$ statistic using $(y_i - \hat{y}_i)^2 / \sigma_i^2$; where $\hat{y}$ is the model estimate of $y_i$ and $\sigma_i$ is the error after binning. The reduced $\chi^2$ is another feature for our classifier (Figure 4).

### 4.3. Flare Finding

X-ray flares are common features in active stars. To test for the existence of flares, we decomposed each light curve into a piecewise constant representation and then looked for segments with elevated count rates compared to adjacent segments. We used the Bayesian blocks technique to construct the piecewise constant segments (Scargle 1998). This technique is designed for astronomical count data with Poissonian noise and is based on the Bayesian formalism. It relies on comparing two hypotheses—the unsegmented hypothesis where the light curve can be described with one rate, and the segmented hypothesis where the light curve is described with two rates. The likelihood that the count rate is constant is given by:

$$L(H_{\text{unseg}}|\text{Data}) = \frac{\Gamma(A + 1)}{(B + 1)^{A+1}},$$

from Equation (29) in Scargle (1998), where $A$ is the number of photons and $B$ is the number of bins. On the other hand, the likelihood of the segmented model is:

$$L(H_{\text{seg}}|\text{Data}) = \frac{\Gamma(A_1 + 1)}{(B_1 + 1)^{A_1+1}} \times \frac{\Gamma(A_2 + 1)}{(B_2 + 1)^{A_2+1}},$$

where $A_1$, $B_1$ and $A_2$, $B_2$ are the number of photons and number of bins in segment one and segment two respectively. To compare the two hypotheses, we calculated the odds ratio:

$$O_{12} = \frac{L(H_{\text{unseg}}|\text{Data})}{L(H_{\text{seg}}|\text{Data})}.$$
Figure 5. Fraction of sources of each type in the training set according to the number of flares found.

(A color version of this figure is available in the online journal.)

Figure 6. Confusion matrix from performing 10 fold cross-validation on the training set using the RF classifier with only time-series features. The color bar represents the true positive rate.

(A color version of this figure is available in the online journal.)

Richards et al. (2011) in the classification of variable stars. These are general statistical measures that do not depend on the time ordering of the measurements, e.g., fractional variability, mean, and standard deviation. Detailed descriptions of these features can be found in Table 4.

4.5. Accuracy on Training Set

To evaluate the accuracy of our classifier, we used the method of cross-fold validation. We divided the training sample into 10 sets, trained with nine sets, used the model created to classify the remaining sample set and then repeated for 10 different combinations. The overall accuracy is the total number of correctly classified samples divided by the total number of samples in the training set. Using only time-series features, the overall accuracy is \(77\%\).

Figure 6 shows the confusion matrix, where the number in each square represents how the detections are classified. The sum of each row of the confusion matrix is the total number of detections in that class. The numbers in the diagonals are detections that have been correctly classified. GRBs, SSSs, and ULXs are the three worst performing classes. This is not unexpected since SSSs and ULXs have no distinguishing time series features. In contrast, stars, XRBs, and CVs performed relatively well and are usually only confused with each other. From Figure 2, it can be seen that XRBs and CVs share semi-periodic temporal behavior while stars have distinguishing flares. It is also worth noting that sources of all types are most likely to be mis-classified as stars. Since a significant proportion of our training set are stars, the classifier optimizes for accuracy by labeling sources for which it does not have sufficient information as the majority class.

Figure 7 shows a plot of the missed detection rate versus the false positive rate, known as the Receiver Operating Characteristic (ROC) plot. Missed detection is \(1 - \text{true positive rate}\), which is the proportion of samples classified as the actual class; false positive rate is the proportion of samples not of the class but classified as such. We created the ROC plot by transforming the results of the multi-class classification into a binary classification, i.e., each sample either belongs to the actual class (true positive) or it does not (false positive). A well performing classifier should be able to provide a low missed detection rate with a low false positive rate, i.e., be on the bottom-left part of the ROC plot. Figure 7 shows CVs are the best performing source type, even though they only achieved an accuracy of 52%. This accuracy is lower than what may be expected since the missed detection rate is only \(10\%\) when the false positive rate is \(20\%\). This is because the test set is unbalanced, a small number of stars mis-classified as CVs would not significantly decrease the accuracy for stars, and therefore would not lead to a high false positive rate.

The R package randomForest also has the ability to calculate relative feature importance. The importance of each feature is estimated by calculating the total decrease in Gini impurity (Equation (1)) from using that feature, averaged over all the trees in the forest. Figure 8 shows the mean decrease in Gini impurity for the time series features. The five most relevant features are, in order of importance: max_slope, powerlaw_goodness_of_fits, median_abs_dev, powerlaw_C, and LombScargle_period1 (all as defined in Table 4).
5. CLASSIFICATION WITH CONTEXTUAL FEATURES

In Section 4, we showed that time-series features have some discriminative power, but that the classification accuracy is insufficient for practical use. In this section, we expand our feature set to include hardness ratios, optical/near infra-red (NIR)/radio cross-matches, proximity to galaxies, and Galactic positions to improve the classification accuracy. We begin by describing each of these features and Table 5 is a summary of the contextual features used in this paper.

5.1. Description of Features

5.1.1. Hardness Ratios

Hardness ratio is a crude proxy for the shape of the X-ray spectrum and it has been used with moderate success to classify X-ray sources (Kahabka et al. 1999). The XMM-Newton EPIC cameras cover the energy band from 0.2 keV to 12.0 keV. The photons gathered are separated into five bands by the 2XMM pipeline, from which four hardness ratios are calculated as follows:

$$HR_n = \left( \frac{R_{n+1} - R_n}{R_{n+1} + R_n} \right)$$

where $R_n$ is the count rate in the $n$th energy band (see Table 5 for the energy range covered by each band). If both bands have count rates within $3\sigma$ of zero, the resulting hardness ratio can be unpredictable as one is essentially dividing one very small number by another very small number. For these cases, we set the hardness ratio to $-10.0$ as a flag.

5.1.2. Optical/NIR Cross-matches

For optical and NIR cross-matching, we used the Naval Observatory Merged Astrometric Dataset (NOMAD; Zacharias et al. 2004). NOMAD is a conglomeration of various optical photometry and astrometry catalogs and the NIR 2MASS catalog.

To estimate the probability of a chance cross-match, we used the Bayesian method from Budavári & Szalay (2008) where we compared the hypothesis that the cross match is genuine to the alternate hypothesis that the source and the optical counterpart are two unrelated sources. The ratio of the likelihood of these two hypotheses is known as the Bayes factor, $\mathcal{B}$, given by the formula:

$$\mathcal{B} = \frac{2}{\psi_1^2 + \psi_2^2} \exp \left( -\frac{\phi^2}{2 (\psi_1^2 + \psi_2^2)} \right),$$

Table 5

| Feature                     | Description                                                                 |
|-----------------------------|-----------------------------------------------------------------------------|
| HR1                         | $(R_2 - R_1)/(R_2 + R_1)$ where $R_1$ and $R_2$ are the count rates in the 0.2–0.5 keV and 0.5–1.0 keV bands, respectively |
| HR2                         | $(R_3 - R_2)/(R_3 + R_2)$ where $R_2$ and $R_3$ are the count rates in the 0.5–1.0 keV and 1.0–2.0 keV bands, respectively |
| HR3                         | $(R_4 - R_3)/(R_4 + R_3)$ where $R_3$ and $R_4$ are the count rates in the 1.0–2.0 keV and 2.0–4.5 keV bands, respectively |
| HR4                         | $(R_5 - R_4)/(R_5 + R_4)$ where $R_4$ and $R_5$ are the count rates in the 2.0–4.5 keV and 4.5–12.0 keV bands, respectively |
| $B$ mag                     | $B$-band magnitude                                                          |
| $V$ mag                     | $V$-band magnitude                                                          |
| $R$ mag                     | $R$-band magnitude                                                          |
| $B - V$                     | $B$-band magnitude minus $V$-band magnitude                                 |
| $H$ mag                     | $H$-band magnitude                                                          |
| $J$ mag                     | $J$-band magnitude                                                          |
| $K$ mag                     | $K$-band magnitude                                                          |
| $J - H$                     | $J$-band magnitude minus $H$-band magnitude                                 |
| $J - K$                     | $J$-band magnitude minus $K$-band magnitude                                 |
| Optical Bayes               | Bayes factor for optical cross-match                                         |
| Radio                       | Bayes factor for radio cross-match                                          |
| isGalaxyAssociation         | Whether there is a galaxy association (Yes or No)                           |
| Luminosity                  | If there is a galaxy association, then calculate the X-ray luminosity of the source using the galaxy’s distance |
| r_ratio                     | If there is a galaxy association, distance to galaxy center divided by radius of the galaxy |
| galAngSep                   | Angular separation between the centroid of a galaxy and the position of the source |
| Gal_lat                     | Galactic latitude                                                          |
| Gal_lon                     | Galactic longitude                                                         |

Figure 8. Relative importance of the time series features. The features are described in detail in Tables 4.
where $\psi_1$ and $\psi_2$ are the resolution of the two catalogs in arcsec and $\phi$ is the angular separation between the two sources. A high Bayes factor favors the hypothesis that the cross-match is genuine. This calculation does not take into account the sky density of the optical sources. For our feature set, we included the $B$, $V$, $J$, $H$, $K$-band magnitudes if a cross match was found, and the corresponding Bayes factor. If no cross-match was found, the magnitude was set to 100 as a null flag.

5.1.3. Radio Cross-matches

We cross-matched the 2XMMi sources with three radio catalogs—the NRAO VLA Sky Survey (NVSS; Condon et al. 1998), the Sydney University Molonglo Sky Survey (SUMSS; Mauch et al. 2003), and the Second Epoch Molonglo Galactic Plane Survey (MGPS-2; Murphy et al. 2007). Together, these catalogs provide all-sky coverage of the radio sky. NVSS was a 1.4 GHz radio survey with the Very Large Array covering the entire sky north of declination $-40^\circ$. SUMSS was the counterpart survey with the Molonglo telescope of the southern sky (south of declination $-30^\circ$) at 843 MHz; MGPS-2 was the Galactic plane radio survey at the same frequency. The positional accuracy of NVSS is $<1''$ for sources stronger than 15 mJy, and 7'' in the survey limit. For SUMSS and MGPS-2, the position accuracy is poorer but typically better than 5''. Since the angular resolution of $XMM$-Newton EPIC is better than those of NVSS or SUMSS, for our cross-matching we used a 3$\sigma$ search radius based on the radio catalogs. We also included the Bayes factor (Equation (9)) to estimate the likelihood of a cross-match. The relatively low sky density of radio sources means that a spurious match is unlikely.

5.1.4. Associations with Galaxies

X-ray sources that correspond to the nuclei of galaxies are likely to be AGN, while non-nuclear extragalactic X-ray sources with luminosities of more than $10^{38}$ erg s$^{-1}$ are potential ULX candidates, but can also be foreground stars, XRBs, CVs, or background AGN. We cross-matched the 2XMMi sources with the Third Reference Catalog (RC3) of galaxies (de Vaucouleurs et al. 1991) to find possible galaxy associations. RC3 contains more than 23,000 galaxies, including almost all galaxies with apparent diameters greater than 1'. RC3 contains information on the galaxy center position, the major and minor diameters of the $D_{25}$ isophote (roughly the domain of the galaxy) as well as the position angle. We determined $\alpha$, the ratio between the angular separation between the source and the galaxy center and the elliptical radius $R_{25}$. If $\alpha < 1.5$, then we considered the source to be associated with the galaxy. For sources associated with a galaxy, we included $\alpha$ and the angular separation in the feature set.

5.1.5. Galactic Coordinates

The last set of features we included is the Galactic position of each source. From Figure 9, it can be seen that XRBs are more likely to cluster along the Galactic plane, while all other source types are distributed isotropically in Galactic coordinates. This motivates the inclusion of Galactic coordinates in the feature set as a way to identify XRBs.

5.2. Accuracy of Training Set

As in Section 4.5, we used 10 fold cross-validation to evaluate the performance of this feature set. Using both the time-series and contextual feature sets, the overall accuracy improved significantly from 77% to 97% with the additional features. Figures 10 and 11 show the confusion matrix and the ROC plot.
6. CATALOG OF PROBABILISTICALLY CLASSIFIED XMM VARIABLE SOURCES

6.1. Results

Using the entire training set, we constructed a RF classification model using the method described in Section 5. Then we applied this classification model to the set of unknown 2XMMi variable sources. For sources where there are more than one detection, we classified each detection separately and combined the results by averaging the output class membership probabilities. Table 6 shows the number of unknown sources classified as one of seven classes. The majority of the unknown sources are classified as stars.

We also compiled a downloadable table of the class membership probabilities. Table 7 shows a portion of that table.

6.2. Evaluation of Results

6.2.1. Comparison with Recent Classifications in Literature

Following the initial source classification (S. Farrell et al., in preparation), a number of sources in the unknown sample have since been classified in the literature. We assessed the accuracy of the classifier by comparing the literature classification to the output of our RF classifier for ∼12% of the unknown sources. Confirming the classification for 411 X-ray sources is beyond the scope of this paper. We found recently confirmed or tentative classifications for 19 sources and they are listed in Table 8. The classifications from our RF classifier agree with the literature classifications in 15 out of 19 cases if we include the two sources that have multiple possible classifications. The misclassifications are due to the source belonging to a novel source type, insufficient information, poor data quality, or problems with the classification in the literature. Of the six misclassifications, three sources have been classified as ULXs by our RF classifier while Kamizasa et al. (2012) regarded them as candidate AGN with immediate-mass black holes based on the presence of X-ray variability. The criteria used by Kamizasa et al. (2012) do not preclude ULXs since they only filtered out sources in known star forming regions and included sources with object type Galaxy shown in the NED databases. All three of the sources classified as ULXs are close to a galaxy in RC3 and have X-ray luminosity of between 10^{39} and 10^{40} erg s^{-1}. Here we briefly discuss three of the other misclassifications.

2XMM J034645.4+680947. This source is classified as an XRB by our classifier but Mak et al. (2011) classified it as a SSS. However, there are a few problems with the literature classification. Mak et al. (2011) only used two hardness ratios in the classification. This is coarser than what we have used, which would have resulted in the loss of information. There are four observations of this source and the hardness ratio only satisfied the criteria for SSS (as defined by Mak et al. 2011) in the two fainter observations. The lack of X-ray flux in the 2–7 keV band could be a selection effect since the hard emission tends to be undetectable in fainter sources. Furthermore, the hardness ratios derived from the 2004 August and 2004 February observations do not classify this source as a SSS. We fitted the 2004 August EPIC spectra that were automatically extracted by the XMM pipeline with a Raymond–Smith model (Raymond & Smith 1977), typical for a SSS. The best-fit parameters are N_H = (0.03 ± 0.03) \times 10^{22} cm^{-2}, kT = (0.79 ± 0.05) keV, and \chi^2/dof = 175.03/183. This is a satisfactory fit, however, the temperature is an order of magnitude higher than typical for a SSS (SSSs peak in the range 20–100 eV; Kahabka & van den Heuvel 2006). From the above arguments, we are skeptical that 2XMM J034645.4+680947 is a SSS. Our RF model classified...
either we lack sufficient information to classify this source as an XRB, SSS, star and ULX with probabilities 0.349, 0.227, 0.222 and 0.16 respectively. This suggests that either we lack sufficient information to classify this source and/or that this source is highly unusual.
Although novae are a subset of CVs we do not have many examples of novae in outburst in the training set, hence to our classifier this is a novel source type. This highlights one of the limitations of supervised classification in that the classifier is incapable of recognizing novel classes.

**2XMM J174016.0−290337.** This source is classified as an XRB by our classifier. Using only X-ray timing and X-ray spectral data, Farrell et al. (2010) identified this source as likely to be a symbiotic XRB, a new and rare sub-class of XRBs composed of a late-type giant accreting matter onto a compact object such as a neutron star. However, with more optical spectral data, Masetti et al. (2012) later identified it as an mCV. There is an optical counterpart in the XMM-Newton error circle with a spectrum that contains strong Balmer, He i, He ii, and Bowen blend emissions, typical of magnetic CVs. Similar to the conclusion made by Farrell et al. (2010), our classifier favors the interpretation of this source as an XRB, giving it a probability of 0.46, but also gives the probability of this being a CV as 0.29. This demonstrates that our classifier is capable of making a conclusion along the same line as an expert in the field using the same information. It is worth noting that this is an unusual source and its X-ray properties do not fit with the interpretation of it being an mCV.

### 6.2.2. Classification of Known Sources in 2XMMi-DR3

Another method we used to evaluate the performance of our classifier is to use the classification model to classify 27 known variable sources in 2XMMi-DR3 that were not in DR2. The 2XMMi-DR3 catalog is an incremental update to the 2XMMi-DR2 catalog and consists of all of 2XMMi-DR2 plus observations made between 2008 August and 2009 October. The 27 sources we have chosen are the targets of observations with known classifications, but which are not in our training set. We classified 22 out of 27 sources correctly, which gives an accuracy of ~81% (Table 9). This is lower than the accuracy from 10-fold cross-validation of ~97%. However, this is not unexpected since the composition of source types in this DR3 subset is vastly different to the training set. For instance, 37% of sources in the DR3 subset are CVs but in our training set, only 4.6% are CVs. In the 10 fold cross-validation, 69% of sources are stars for which the classification accuracy is 99.8%. We were able to classify all seven stars in the DR3 subset correctly. Of the seven DR3 sources that we misclassified, two are unusual sources, 2XMMi J050106.5+451634 is a known magnetar (Rea et al. 2009), a source type that is not in our training set. Although there were magnetars in the variable samples, we excluded them from the training set since there were only very few samples. The other source, 2XMMi J080407.8−450329, is a recurrent supergiant fast X-ray transient (SFXT; Leyder et al. 2007). SFXTs have only been identified recently as a new class of XRBs and are believed to consist of a wind-accreting compact object and an OB super-giant donor star. In both cases, our classifier was not able to devise a correct classification because the correct class is not one that the classifier has knowledge of.

There are two GRBs in our DR3 subset and we correctly classified both of them, despite GRB being a minority class. We repeated the experiment and trained a RF classifier without resampling the data set and found that it was only able to identify one of the two GRBs. This demonstrates that resampling is important for achieving good performance on minority classes.

### 7. ANOMALOUS SOURCES

In the previous section, we used mislabeled instances to highlight one of the issues with supervised classification—namely
that it cannot label novel source types. Uncovering novel but rare source types is a stated goal of many large surveys. In machine learning, this task is known as anomaly detection (Chandola et al. 2009). Anomalies are cases whose proximity to other cases of the same type is small. In the R package randomForest, there is a function to calculate an outlier measure based on the proximity matrix. The proximity matrix Prox is an $n \times n$ matrix (assuming the training set has $n$ cases) where $Prox(i, j)$ is incremented by one if case $i$ and case $j$ both end up in the same terminal node of a tree. $Prox$ is normalized by dividing by the number of trees in the forest. The outlier measure calculates the proximity of the case $i$ to other cases of the same class, using the equation:

$$O(i) = \frac{1}{MAD} \left( \frac{n}{\sum_{j=1}^{n} Prox^2(i, j) - M} \right), \quad (10)$$

where $k$ is the class of case $i$, and $n$ is the number of instances of the class $k$. $O(i)$ is normalized by subtracting $M$, the median of the unnormalized outlier measures, and dividing by the median absolute deviation (MAD). Higher outlier measures mean the source is more anomalous while a low outlier measure means the source is similar to other sources of the same class.

Figure 13 is a plot of the classification margin against the outlier measure for the 408 unknown sources in our test set, excluding the three sources with recent spectroscopic identifications listed in Table 8. The sources marked as red triangles either have outlier measures greater than 10 and/or classification margins of less than −0.3, making them likely to be true outliers. The cut-off is arbitrary and we use it to select a manageable number of potential outlier sources to verify this technique. 12 sources satisfy these criteria and are listed in Table 10. We now discuss possible reasons why these sources have been deemed anomalous.

A common reason for sources to be classified as anomalous is bad data quality. For sources 3, 5, 6, and 12, the source of interest is either very close to an extremely bright source, or within the confines of a diffuse source. These situations can lead to the contamination of the X-ray spectrum (thereby giving unreliable fluxes and HRs) and/or time series. Sources with low signal-to-noise ratios (S/Ns) can also be erroneously classified as anomalous. Source 3 has S/N of less than 10 and we cannot trust the classifier’s determination that it is anomalous. Low S/N means that the hardness ratios will have large error bars and that errors on features will not be used properly in the classification process. Incorporating error bars into the classification algorithm is an area to be addressed in future work.

One of the anomalous sources on our list is source 2, 2XMM J034645.4+680947, which we have already discussed in Section 6.2.1. We will now briefly consider the nature of source 12.

2XMM J180658.7−500250 (source 12). This is the most unusual source in our list. It has a counterpart in 2MASS with $J = 14.144$ mag, $H = 13.446$ mag, and $K = 12.715$ mag, but no other optical or radio counterparts were found in the literature.7 There is also a mid-IR match from the Wide-field Infrared Survey Explorer (WISE) survey with magnitudes at 12 and 22 μm of 7.356 and 5.165, respectively. The mid-IR WISE colors are consistent with a spiral galaxy (Wright et al. 2010).

---

Table 10
Anomalous Unknown Sources

| Number | 2XMM Name                  | Our Classification | Margin | Outlier | Sum Flag | S/R | Notes                                      |
|--------|---------------------------|--------------------|--------|---------|----------|-----|--------------------------------------------|
| 1      | J013612.5+154957          | AGN                | −0.42  | 4.4     | 1        | 23  | AGN candidate (Kamizasa et al. 2012)      |
| 2      | J034645.4+680947          | XRB                | −0.32  | 6.4     | 0        | 31  | Anomalous source, see Section 6.1          |
| 3      | J045445.5−180641          | ULX                | −0.30  | 19.5    | 3        | 7   | Low S/N                                    |
| 4      | J120143.6−184857          | ULX                | −0.27  | 13.6    | 0        | 22  | AGN candidate (Kamizasa et al. 2012)      |
| 5      | J122543.2+332353          | ULX                | −0.30  | 20.1    | 4        | 14  | Next to a bright source                    |
| 6      | J122549.0+333202          | ULX                | −0.18  | 16.7    | 3        | 23  | Next to a bright source                    |
| 7      | J123316.0+000512          | ULX                | −0.40  | 14.4    | 0        | 18  | AGN candidate (Kamizasa et al. 2012)      |
| 8      | J151013.0−034749          | CV                 | −0.16  | 132.1   | 0        | 20  | Large flare                                |
| 9      | J161741.9−833751          | AGN                | −0.32  | 2.4     | 1        | 51  | Stochastic X-ray variability              |
| 10     | J180658.7−500250          | AGN                | −0.49  | 2.6     | 3        | 211 | Anomalous source, Section 7.1              |
| 11     | J181330.6−333627          | CV                 | −0.20  | 416.4   | 3        | 49  | V2694 pulsating star                      |
| 12     | J231818.7−422237          | ULX                | −0.29  | 14.86   | 1        | 32  | Next to a bright source                    |

Notes. Column definitions from the left: Source number; 2XMM name; classification given to the source by our RF classifier; classification margin; outlier measure (Equation (10)); sum flag which is a measure of data quality from the XMM data processing pipeline (scale of 0–4; 0 means good, 4 means source is possibly spurious); signal-to-noise ratio; notes.
indicating that this object may be extragalactic. From the WISE image this source looks unresolved, which gives it an upper limit angular size of $10''$, and if we assume the size of a spiral galaxy to be at least 5 kpc, then we can constrain the distance to this source to be at least 100 Mpc.

We attempted to fit the X-ray spectra with different simple one-component models (power law, blackbody, disk blackbody, thermal plasma, and bremsstrahlung models). The absorbed power-law model produced the best fit with $\chi^2$/dof $= 1888.43/1417$. However, as this is not a statistically acceptable fit we tried more complex two component models (power law + disk blackbody, power law + blackbody, mekal + mekal, bremsstrahlung + blackbody). The two component model that provided the best physical fit to the data is the absorbed power law + disk blackbody model (Figure 13). The best-fit parameters are: $n_H = 2.9 \times 10^{20}$ atoms cm$^{-2}$, $\Gamma = 1.8$, $T = 0.36$ keV. However, it is still not a good statistical fit, with $\chi^2$/dof $= 1702.88/1415$ and significant residuals around 0.5 keV (see Figure 14).

Such low-energy features are reminiscent of those seen in the spectra of the well known ULX Holmberg II X-1. Using high-quality spectra obtained with the XMM-Newton EPIC and Reflection Grating Spectrometer instruments, Goad et al. (2006) detected evidence for a complex of emission lines in the spectra of Holmberg II X-1 with energies between 563 and 577 eV, possibly associated with the O vii triplet. We therefore added a Gaussian component to our best fit absorbed power law plus disk black body model, finding that a broad emission line ($E_{\text{Gauss}} = 0.50 \pm 0.01$ keV, $\sigma = 0.05 \pm 0.01$ keV, equivalent width $= 0.044$ keV, and unabsorbed bolometric flux $= (1.5^{+0.2}_{-0.4}) \times 10^{-13}$ erg cm$^{-2}$ s$^{-1}$) improved the fit significantly to a statistically acceptable $\chi^2$/dof $= 1530.77/1412$. In the case of Holmberg II X-1, Goad et al. (2006) found evidence for the presence of two narrow O vii lines: the forbidden line at 563 eV and the resonance line at 577 eV. However, replacing our broad line with two narrow lines worsened our fit. Goad et al. (2006) also considered the possibility that the emission features they detected were due to the presence of an optically thin thermal plasma. We thus tried refitting our spectra of 2XMM J180658.7−500250 with the MEKAL thermal plasma model replacing the disk blackbody and Gaussian emission features. However, this did not improve the fit ($\chi^2$/dof $= 1678.46/1414$).

The possible coincidence with a spiral galaxy at a redshift between $z \sim 0.01–0.05$ combined with the spectrum reminiscent of Holmberg II X-1 raises the possibility that 2XMM J180658.7−500250 may also be a ULX. If this is the case, the absorbed 0.2−10 keV flux of $(6.05^{+0.06}_{-0.05}) \times 10^{-14}$ erg cm$^{-2}$ s$^{-1}$ using the best-fit power law plus disk black body plus Gaussian model implies a luminosity between $\sim 10^{42}–10^{43}$ erg s$^{-1}$, which would make 2XMM J180658.7−500250 even more luminous than the brightest ULX (and strongest intermediate mass black hole candidate) ESO 243-49 HLX-1 (Farrell et al. 2009). Such luminosities are extremely difficult to explain with a stellar mass black hole, implying a black hole mass of $\gg 1000 M_\odot$. A luminosity this high is much more reminiscent of those observed from AGN, however the disk blackbody temperature is too high for an accretion disk around a supermassive black hole. However, such a disk blackbody temperature is typical of ULXs (Berghera et al. 2008).

In addition, the light curve of this source is also interesting and has a tantalizing hint of periodicity (Figure 15). Periodic variability does not fit with the AGN interpretation and instead favors the compact binary object classification. In addition, the mid-IR color is highly unusual for an AGN, and is instead much more reminiscent of a non-active galaxy. We therefore speculate that 2XMM J180658.7−500250 may be a new member of the extremely rare class of hyper-luminous X-ray sources (HLXs), i.e., ULXs with luminosities in excess of $10^{41}$ erg s$^{-1}$, that potentially represent the best candidates for intermediate mass black holes.

In summary, 2XMM J180658.7−500250 appears to be unusual source that our classifier has rightly picked out as anomalous. More work, such as fitting more complex X-ray spectral models as well as multi-wavelength follow-up observations, is needed to verify its nature.

8. CONCLUSIONS

In this paper, we have tested the performance of the RF classifier with the 2XMmi-DR2 data set. On a 7 class data set with only time series features, we were able to attain a 10 fold validation accuracy of $\sim 77\%$. Time series features do have some
discriminative power, but in the absence of other information, they do not result in a high performing classifier. When we added in contextual features such as hardness ratios, optical/IR/radio cross-matches, Galactic coordinates and proximity to nearby galaxies, the classification accuracy increased to ~97%. This shows that the RF classifier can be a high performing classifier, but only by combining both time-series and contextual features. The same conclusion was made by Palaversa et al. (2013) in their work on the automatic classification of optical stars, in which they found that using both light curve features and colors allowed them to achieve accuracy of 92%. A potential recommendation from our work is that the classifiers for future synoptic variable surveys will need more than just temporal flux measurements to achieve good performance.

We demonstrated the scientific potential of an automatic classifier by applying our RF classifier to 411 unknown variable sources. To test the reliability of such automatic classification, we found recent classifications in the literature for 19 sources and checked the literature’s suggested classification against the output from our classifier. Our classification agrees with the literature in 13 out of the 19 sources (accuracy of 68%). The mislabeled cases are due to a source belonging to a new and unseen class or because the classification made in literature used information (such as optical spectra) that were not available to us. We also used our RF classifier on a known subset of target sources in 2XMM-DR3. We were able to classify 22 out of 27 sources correctly (accuracy of 81%). The mislabeled sources are again of unknown source types, or are unusual members of one of the known source types.

In the DR3 verification exercise, we showed that the RF classifier can accurately classify GRBs, a heavily under-represented class. This performance was achieved by oversampling the minority classes.

To find anomalous sources, we used the classification margin and the outlier measure from the RF package. Most of the high potential anomalous sources we found contained data quality issues. One source in our list did look genuinely unusual (2XMM J180658.7–000250) and further work needs to be done to determine its true nature.

There are two areas for improvement on the algorithm front. First, to the best of our knowledge, current machine learning algorithms (including RF) do not take into account the error bars in the features. In astronomy, accurate measurement errors are readily available and provide valuable information, and should be incorporated into the machine learning algorithm. One simple way to do this is to apply a weighting to reflect the size of the error. This needs to be done in such a way that would propagate the error to the classification accuracy. Second, the RF classifier lacks interpretability. For an individual source, the RF classifier does not allow the user to pinpoint the feature which led to the classification, which is something that a human expert can easily provide. However, RF can provide a measure of feature importance measured using all the samples in the training set.

Automatic classification will likely play a major role in future synoptic surveys across all wavelengths. In this paper, we have shown that the RF classifier can achieve excellent performance. We envision that a similar model can be built into the pipeline for time-domain surveys on the SKA and the LSST, where the goal will be to produce probabilistic classifications as a value-added component to the catalogs.

This research was conducted by the Australian Research Council Centre of Excellence for All-sky Astrophysics (CAASTRO), through project number CE110001020. K.K.L. is supported by a university postgraduate award from the University of Sydney and a scholarship from CAASTRO. S. A. Farrell is the recipient of an Australian Research Council Post Doctoral Fellowship, funded by grant DP110102889. Based on observations from XMM-Newton, an ESA science mission with instruments and contributions directly funded by ESA Member States and NASA. This work made use of the 2XM Serendipitous Source Catalogue, constructed by the XMM-Newton Survey Science Centre on behalf of ESA. This research has also made use of the NASA/IPAC Extragalactic Database (NED) which is operated by the Jet Propulsion Laboratory, California Institute of Technology, under contract with the National Aeronautics and Space Administration, and the SIMBAD database, operated at CDS, Strasbourg, France. We thank the referee for comments that have helped to improve the paper.

REFERENCES

Atlee, D. W., & Gould, A. 2007, ApJ, 664, 53
Benz, A. O., & Güdel, M. 2010, ARA&A, 48, 241
Bergera, C. T., Weaver, K. A., Colbert, E. J. M., & Roberts, T. P. 2008, ApJ, 687, 471
Bloom, J. S., Richards, J. W., Nugent, P. E., et al. 2012, PASP, 124, 1175
Brodie, J. 2017, Adv. Learn. 45, 5
Budavári, T., & Szalay, A. S. 2008, ApJ, 679, 301
Cardiles, S., Budavári, T., Heinis, S., Priève, C., & Szalay, A. S. 2010, ApJ, 712, 511
Caruana, R., & Niculescu-mizil, A. 2006, in Proceeding of the 23rd Int. Conf. on Machine Learning (ICML06). ACM, New York, NY, 161
Chandola, V., Banerjee, A., & Kumar, V. 2009, ACM Comput. Surv., 41, 15
Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. 2002, J. Artif. Int. Res., 16, 321
Condon, J. J., Cotton, W. D., Greisen, E. W., et al. 1998, AJ, 115, 1693
Cordes, J. M., Lazio, T. J. W., & McLaughlin, M. A. 2004, NewAR, 48, 1459
de Vaucouleurs, G., de Vaucouleurs, A., Corwin, H. G., Jr., et al. 1991, Third Reference Catalogue of Bright Galaxies. Volume I: Explanations and References. Volume II: Data for Galaxies between 0h and 12h. Volume III: Data for Galaxies between 12h and 24h (New York: Springer)
Djorgovski, S. G., Mahabal, A. A., Donalek, C. et al. 2012, in Proceedings of the IEEE eScience 2012 Conference (IEEE Press), 6404437
Dubath, P., Rimoldini, L., Stévèges, M., et al. 2011, MNRAS, 414, 2602
Duda, R., Hart, P., & Stork, D. 2001, Pattern Classification, Pattern Classification and Scene Analysis: Pattern Classification (2nd ed.; New York: Wiley)
Farrell, S. A., Gosling, A. J., Webb, N. A., et al. 2010, A&A, 523, 580
Farrell, S. A., Webb, N. A., Barret, D., Godet, O., & Rodrigues, J. M. 2009, Natur, 460, 73
Goad, M. R., Reeves, T. P., Reeves, J. N., & Utley, P. 2006, MNRAS, 365, 191
Hearn, D. R., & Richardson, J. A. 1977, ApJl, 213, L115
Heinke, C. O., Tombisk, J. A., Yusef-Zadeh, F., & Grindlay, J. E. 2009, MNRAS, 400, 1627
Hofmann, F., Pietsch, W., Henze, M., et al. 2013, A&A, 555, A65
Hui, C. Y., Srim, K., & Choi, C.-S. 2012, MNRAS, 421, 341
Israel, G. L., Hummel, W., Covino, S., et al. 2002, A&A, 386, L13
Kahabka, P., Pietsch, W., Filipović, M. D., & Haberl, F. 1999, A&AS, 136, 81
Kahabka, P., & van den Heuvel, E. P. J. 2006, in Compact Stellar X-Ray Sources, ed. W. Lewin & M. van der Klis (Cambridge Astrophysics Series, No. 39; Cambridge: Cambridge Univ. Press), 461
Kanizsay, A., Terasawa, Y., & Nakajima, H. 2012, ApJ, 751, 39
Leyer, J.-C., Walter, R., Lazaros, M., Masetti, N., & Produit, N. 2007, A&A, 465, L35
Liaw, A., & Wiener, M. 2002, R News, 2, 18
Lin, D., Webb, N. A., & Barret, D. 2012, ApJ, 756, 27
Lomb, N. R. 1976, Ap&SS, 39, 447
Longair, M. 2011, High Energy Astrophysics (Cambridge University Press)
Mak, D. S. Y., Pun, C. S. J., & Kong, A. K. H. 2011, ApJ, 728, 10
Masetti, N., Nucita, A. A., & Parisi, P. 2012, A&A, 544, A114
Mattijevich, G., Prša, A., Orosz, J. A., et al. 2012, AJ, 143, 123
Mutsch, T., Murphy, T., Butler, H. J., et al. 2003, MNRAS, 342, 1117
McGlynn, T. A., Suchkov, A. A., Winter, E. L., et al. 2004, ApJ, 616, 1284
Merloni, A., Pedech, P., Becker, W., et al. 2012, arXxiv:1209.3114
Murphy, T., Chatterjee, S., Kaplan, D. L., et al. 2013, PASA, 30, 6
