Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **517,** 3362–3376 (2022) https://doi.org/10.1093/mnras/stac2760
Advance Access publication 2022 October 3

Machine learning-based search for cataclysmic variables within *Gaia* Science Alerts

D. Mistry [1]★ C. M. Copperwheat,[1] M. J. Darnley [1] and I. Olier [2]

[1]*Astrophysics Research Institute, Liverpool John Moores University, IC2, Liverpool Science Park, 146 Brownlow Hill, Liverpool L3 5RF, UK*
[2]*School of Computer Science and Mathematics, Liverpool John Moores University, James Parsons Building, 3 Byrom Street, Liverpool L3 3AF, UK*

Accepted 2022 September 14. Received 2022 August 31; in original form 2022 June 15

ABSTRACT

Wide-field time domain facilities detect transient events in large numbers through difference imaging. For example, Zwicky Transient Facility produces alerts for hundreds of thousands of transient events per night, a rate set to be dwarfed by the upcoming Vera C. Rubin Observatory. The automation provided by machine learning (ML) is therefore necessary to classify these events and select the most interesting sources for follow-up observations. Cataclysmic variables (CVs) are a transient class that are numerous, bright, and nearby, providing excellent laboratories for the study of accretion and binary evolution. Here we focus on our use of ML to identify CVs from photometric data of transient sources published by the *Gaia* Science Alerts (GSA) program – a large, easily accessible resource, not fully explored with ML. Use of light-curve feature extraction techniques and source metadata from the *Gaia* survey resulted in a random forest model capable of distinguishing CVs from supernovae, active galactic nuclei, and young stellar objects with a 92 per cent precision score and an 85 per cent hit rate. Of 13 280 sources within GSA without an assigned transient classification our model predicts the CV class for ∼2800. Spectroscopic observations are underway to classify a statistically significant sample of these targets to validate the performance of the model. This work puts us on a path towards the classification of rare CV subtypes from future wide-field surveys such as the Legacy Survey of Space and Time.

Key words: cataclysmic variables – surveys – methods: data analysis – techniques: photometric, spectroscopic.

1 INTRODUCTION

Over the last few decades, the increasing depth and breadth of imaging produced by wide-field survey facilities has brought forth a revolution in time domain astronomy. Their ability to rapidly and repeatedly image huge areas of the sky combined with the technique of difference imaging, in which the new sky image is subtracted from a reference image, has greatly increased the discovery potential of time varying sources (Kulkarni 2020). Surveys include the Catalina Real-time Transient Survey (CRTS; Drake et al. 2009), Palomar Transient Factory (PTF; Law et al. 2009), Panoramic Survey Telescope and Rapid Response System (Pan-STARRS; Morgan et al. 2012), Zwicky Transient Facility (ZTF; Bellm et al. 2019), and All-Sky Automated Survey for SuperNovae (ASAS-SN; Kochanek et al. 2017). In addition, the space-based *Gaia* mission (Gaia Collaboration et al. 2016b), primarily purposed for astrometry, has been recognized as a powerful tool for time domain astronomy.

Difference imaging can reveal genuine astrophysical sources that have changed in brightness or position and artefacts posing as such. Artefacts can make up a significant proportion of events found from difference imaging, these consist of poorly subtracted galaxies, cosmic rays, point spread function (PSF) haloes, defective pixels, and CCD edge effects (Goldstein et al. 2015). The vetting of such artefacts is no longer solely performed by human inspection, but is now reliant in large part on automated pipelines (e.g., Cao, Nugent & Kasliwal 2016; Mahabal et al. 2019). The classification of astrophysical sources has historically been performed by human inspection (e.g., Zwicky 1964; Strolger et al. 2004; Rest et al. 2014), however, technological advancements in the design of telescopes, detectors, and computing over recent decades have led to a rapid rise in transient events to classify, thus fuelling the desire for ever more efficient classification methods.

Currently ZTF produces up to 10^6 alerts (individual photometric data points of time variable sources) per night, and in a few years' time the Vera C. Rubin Observatory Legacy Survey of Space and Time (LSST; Ivezic et al. 2019) will be generating up to 10^7 alerts per night (Matheson et al. 2021). Only a small fraction of associated transient sources can benefit from follow-up observations due to the limited availability of dedicated facilities and the associated telescope time constraints. As a consequence, one must be selective about the sources that are followed up, these will be those which provide the greatest potential in furthering our understanding of the transient classes to which they belong. Finding these sources requires some level of initial source classification to distil the incoming alert stream to a list of such targets. An additional requirement is the need for prompt classification as early time follow-up can be highly informative, for example, spectroscopy obtained within a few days of a supernova (SN) explosion can inform different progenitor models (Khazov et al. 2016). To address these challenges, real time transient event processing must be accommodated into survey architecture. The level of automation needed can be achieved by integration of

★ E-mail: d.mistry@2018.ljmu.ac.uk

machine learning (ML). Implementation of ML requires little or no human intervention beyond the training of the associated algorithms with labelled data. The resultant models are capable of rapidly classifying unseen examples, requiring only a few seconds at most to complete each task.

Survey facilities such as ZTF have adopted ML to perform tasks such as separating out real transient events from artefacts (bogus or false positives; Cao et al. 2016); and to separate spatially extended targets (such as galaxies) from stars (Tachibana & Miller 2018). This use of ML helps to pare down the large influx of alerts generated by difference imaging. Alerts brokers have been used by ZTF to ingest and characterize the nature of their alerts, serving them to the astronomical community. One such broker is the Automatic Learning for the Rapid Classification of Events (ALeRCE; Förster et al. 2021). The AleRCE pipeline makes use of science, reference and difference images for rapid classification of events (Carrasco-Davis et al. 2021), and multiband light curves for longer term characterization (Sánchez-Sáez et al. 2021). The pipeline uses these inputs to group events into several transient classes, such as cataclysmic variables (CVs), SNe, active galactic nuclei (AGN), variable stars, and young stellar objects (YSOs). Aside from use within brokers, ML has been used for classification of SN subclasses (Gabruseva, Zlobin & Wang 2020; Fremling et al. 2021) and galaxy morphology classification (Dieleman, Willett & Dambre 2015).

An area of time domain where ML classification will be of key importance is that of CVs (extensively covered in Warner 1995; Hellier 2001), systems that are used to develop our understanding of binary evolution (e.g. Kato & Hachisu 2012; Pala et al. 2022; van Roestel et al. 2022). These are semi-detached binary systems composed of a white dwarf (WD) and a late-type main-sequence companion. The companion (or donor) is Roche lobe filling, leading to a transfer of mass to the WD via the inner Lagrangian point. In the majority of cases this results in the formation of an accretion disc. However, where the WD is strongly magnetic the accretion mechanism changes. In polars (or AM Her stars; Cropper 1990) the strong magnetic field ($\sim 10 \leq B \leq 80$ MG) causes the accretion flow from the donor to be funnelled by field lines on to one or both of the WD magnetic poles. In intermediate polars (Patterson 1994) ($\sim 1 \leq B \leq 10$ MG) a partial accretion disc may form where only the inner regions of the disc are magnetically disrupted. CVs give rise to a plethora of observable phenomena. Examples include the highly energetic eruptions of classical novae (CNe), caused by the violent expulsion of the accreted shell of matter from the surface of the WD driven by runaway thermonuclear reactions (Starrfield, Iliadis & Hix 2016); the less energetic but much more frequent outbursts of dwarf novae (DNe) modelled by thermal/viscous instabilities in the accretion disc (Osaki 1996); and the recently identified micronovae, also believed to be thermonuclear runaway events, though localized to magnetically confined regions on the WD surface (Scaringi et al. 2022a,b). Large area time domain surveys are more likely to catch these events in action by monitoring larger numbers of sources.

These systems aid our understanding of the currently uncertain progenitor scenarios of Type Ia SNe (Jha, Maguire & Sullivan 2019), where both single and double degenerate pathways exist. Novae such as M31N 2008-12a (Darnley & Henze 2020) are the most promising single degenerate pathway, these comprise a near Chandrasekhar mass WD accreting at high rates. A double degenerate pathway may be provided by the ultrashort period (5–65 min) helium-rich CVs that make up the AM CVn subclass (Solheim 2010). CVs with known orbital periods are also important for such binary evolution studies, these enable the masses and radii of the individual stars to be determined. The orbital period can be deduced in the SU UMa DN subtypes from an eruption-induced periodic 'superhump' variation in the light curve (Patterson et al. 2005); from radial velocity variations in spectral lines (Inight et al. 2022); or in eclipsing systems via the periodic occultation of the WD and accretion disc by the donor (Copperwheat et al. 2010). The exploration of wide-field survey photometry with ML techniques provides an avenue for the discovery of many more of these transients. This is an active area of research, where examples include work by Neira et al. (2020) to distinguish between CVs, SNe, and several other classes from a data set constructed from CRTS light curves; and the classification of alerts from the ZTF stream within the AleRCE alerts broker pipeline (Sánchez-Sáez et al. 2021).

In this work, we describe our exploration of data generated by the *Gaia* spacecraft (Gaia Collaboration et al. 2016a) to identify new members of the CV population. *Gaia* is now recognized as a powerful tool for transient detection, with *Gaia* Science Alerts (GSA; Hodgkin et al. 2021) providing alerts of newly discovered transient sources at a current rate of ~ 12 d^{-1} by repeatedly scanning the whole sky. The cadence of associated light curves is dictated by the '*Gaia* scanning law' (Gaia Collaboration et al. 2016a) – typically, a pair of observations separated by 106.5 min is separated by another pair 2–4 weeks later. The photometry is precise to 1 per cent at $G = 13$, and 3 per cent at $G = 19$. This resource therefore provides a stable platform from which to evaluate ML-based classification. In Section 2, we describe the classified transients of GSA, the methods used to extract relevant descriptive characteristics from their light curves, and the additional metadata gathered from the survey for each source. In Section 3, we describe how the resultant data set was used to train several ML algorithms to perform a set of classification tasks, along with a description of how the resultant models can be evaluated. In Section 4, we detail the performance of each algorithm. Finally, we discuss the outcomes of our exploration of GSA along with a description of a pilot study involving spectroscopic classification to validate predictions made by our best performing model (Section 5).

2 DATA SET

2.1 *Gaia* alerts and EDR3

As of 2021 June, close to 18 000 transient sources had been listed within the *Gaia* transient alerts stream;[1] just over 4700 of which had been assigned class labels. The classifications are based upon human inspection of *Gaia* data in combination with the results of positional cross-matching with the SIMBAD (Wenger et al. 2000), the NASA/IPAC Extragalactic Database (NED), and the International Variable Star Index (VSX) data bases, and YSO catalogues (see section 2.7.7 of Hodgkin et al. 2021) to identify already-confirmed transient or variable objects. This information is aided by the hourly parsing of 27 major transient survey websites for reported discoveries that also contain classification information, these include Transient Name Server (TNS),[2] CRTS, ASAS-SN, and Astronomer's Telegrams.[3] Further details regarding the alerts filtering and classification process are contained in Hodgkin et al. (2021).

The process of training and validating ML models requires accurate class labels. Whilst we believe the aforementioned process of class assignment can reliably provide this accuracy, an inspection of class labels for a sample of these sources was performed for a

[1] http://gsaweb.ast.cam.ac.uk/alerts/alertsindex
[2] https://www.wis-tns.org/
[3] https://www.astronomerstelegram.org

level of verification. Of the 2713 SNe, 2530 are spectroscopically confirmed according to TNS, Astronomer's Telegrams contain details of spectroscopic classification for the remainder. Of the 613 *Gaia* labelled CVs, 471 are associated with known/confirmed CVs according to the comments associated with the *Gaia* classifications. Comparison with VSX confirms this along with either a confirmation of CV status or candidate status for the remainder through references to relevant research papers and Astronomer's Telegrams. *Gaia*'s comments associated with sources labelled as AGN and YSO show 929 of the 940 transients labelled as AGN, and 184 of the 190 transients labelled as YSOs are associated with known/confirmed AGN and YSOs, respectively. This was verified for a sample of these sources by examining records within TNS and associated links (e.g. SIMBAD). The remaining candidate AGN and YSOs were not further considered for this work.

Our data set is composed of features extracted from light curves of these classified targets within *Gaia*'s alert stream along with their associated class labels. Supplementary data for these targets may be available within the data base of *Gaia* Early Data Release 3 (EDR3; Lindegren et al. 2021; Riello et al. 2021) in the form of astrometric and further photometric data such as parallax, proper motion, and photometric colour provided by the low-resolution photometry ($R = 100$) of blue and red photometers onboard *Gaia*. A coordinate cross-match with EDR3 provides this metadata for ∼45 per cent of sources within our data set. This metadata has also been incorporated as a set of supplementary features.

Of the 4697 classified targets incorporated into our data set, SNe account for 58 per cent of classified targets, AGN make up 21 per cent, CVs and YSOs constitute 13 per cent and 3 per cent, respectively, while microlensing, tidal disruption events, and various other classes account for the remainder.

The majority of GSA classifications come from dedicated spectroscopic follow-up programs such as Public ESO Spectroscopic Survey of Transient Objects (PESSTO; Smartt et al. 2015) and Spectral Energy Distribution Machine (SEDM; Blagorodnova et al. 2018) that are heavily biased towards SN classification. The class fractions of classified targets are generally dictated by what has been chosen to be classified, with unusual or ambiguous examples often overlooked, and therefore it must be noted that these fractions may not be representative of the entire sample of GSA targets.

2.2 Light-curve feature extraction

From source light curves we extracted quantitative characteristics (or features) that describe their variability. These composed of simple statistical and periodicity-based features in Table 1 along with features obtainable from the Feature Extractor for Time Series (FEETS) package (Cabral et al. 2018) based on the light-curve data available (magnitude and time, no error measurements), a selection of which are shown in Table 2. They consist of statistical, periodicity, and percentile-based features.

2.3 Supplementary features

Supplementary data (or metadata) from *Gaia* EDR3 relating to position, photometry, and astrometry are incorporated as data set features. Positional features consist of: right ascension, declination, Galactic (and ecliptic) longitude and latitude, along with associated errors. Photometric features encompass the mean flux from the red and blue photometers (BP and RP) and that from *G*-band photometry, the associated mean magnitudes, colours (BP − RP, BP − *G*, *G* − RP), and associated errors. Proper motion and parallax (along with their errors) are included as astrometric features. A full list is displayed in Table 3, while further details are available within the *Gaia* EDR3 documentation.[4]

Table 1. Features extracted from light curves (without FEETS package).

Feature	Description
mean_mag	Mean of magnitudes
median_mag	Median of magnitudes
std_mag	Standard deviation of magnitudes
mad_mag	Median absolute deviation of magnitudes
min_mag	Minimum magnitude (maximum brightness)
max_mag	Maximum magnitude (minimum brightness)
n_obs	Number of observations
diff_min_mean	Difference between *min_mag* and *mean_mag*
diff_min_median	Difference between *min_mag* and *median_mag*
detected_time_diff	Time span of observations
n_peaks_rm_x_y	Number of observations within a rolling window of y observations that are brighter than x magnitudes of the median magnitude of that window ($x = 1, 2, 3, 4, or\ 5$, $y = 7$).
kurtosis	Kurtosis of the magnitudes
skew	Skewness of the magnitudes
pwr_max	Largest power value in the Lomb–Scargle periodogram
freq_pwr_max	Frequency corresponding to *pwr_max*
FalseAlarm_prob	Estimate of the false alarm probability given the height of the largest peak in the periodogram (see https://docs.astropy.org/en/stable/api/astropy.timeseries.LombScargle.html#astropy.timeseries.LombScargle.false_alarm_probability)

3 METHOD

3.1 Machine learning algorithms

The data set described above can be used to evaluate the ability of machine learning (ML) algorithms to identify CVs within GSA. The algorithms whose performance we evaluated are scikit-learn's (Pedregosa et al. 2011) PYTHON implementation of random forest (RF; Breiman 2001), AdaBoost (ADB; Freund & Schapire 1997), k-nearest neighbours (KNNs; Zhang 2016), and support vector machines (SVMs; Cortes & Vapnik 1995). Also used are the extreme gradient boosting (XGBoost) algorithm (Chen & Guestrin 2016) and Keras (Chollet et al. 2015) implementation of an artificial neural network (ANN) in the form of a multilayer perceptron (MLP) – a fully connected multilayer ANN (Kruse et al. 2022).

RF, ADB, and XGBoost are implementations of an ensemble of decision trees (Rokach & Maimon 2008). Based on the features provided, decision trees perform successive binary splits of the training data set in a way that resultant groups are as different from one another as possible, and closer to a homogeneity of class. The resultant model uses this tree structure to classify unseen examples. RF classifies based on a voting system using the predictions of a random collection of decision trees, the class with the most votes is our model's prediction. Each tree is trained on a modified version of the original training set (bootstrap aggregation) and a random subset of features to introduce uncorrelated trees. ADB (Freund & Schapire 1997) combines decision trees sequentially, weights are

[4] https://gea.esac.esa.int/archive/documentation/GEDR3/Gaia_archive/chap_datamodel/sec_dm_main_tables/ssec_dm_gaia_source.html

Table 2. A small selection of features available from the FEETS package. The full list is available at https://feets.readthedocs.io/en/latest/tutorial.html along with detailed explanations. Of the full list, only those requiring a magnitude and time, or just magnitude data, were implemented here.

Feature	Description
Amplitude	Half of the difference between the median of the maximum 5 per cent and the median of the minimum 5 per cent magnitudes
AndersonDarling	The Anderson–Darling test is a statistical test of whether a given sample of data is drawn from a given probability distribution (normal distribution)
Autocor_length	Cross-correlation of a signal with itself
Eta_e (η^e)	Variability index η is the ratio of the mean of the square of successive differences to the variance of data points
*FluxPercentileRatioMid***X**	Ratio of centred flux percentile ranges. If $F_{5,95}$ is the difference between the 95th and 5th percentile of ordered magnitudes, then *FluxPercentileRatioMid***X** $= F_{40,60}/F_{5,95}$, $F_{32.5,67.5}/F_{5,95}$, $F_{25,75}/F_{5,95}$, $F_{17.5,82.5}/F_{5,95}$, and $F_{10,90}/F_{5,95}$, for **X** = 20, 35, 50, 65, and 80, respectively
*Freq***i***_harmonics_amplitude_j*	Amplitude of the *j*th harmonic of the ***i***th frequency component of the Lomb–Scargle periodogram
Gskew	Median-of-magnitudes-based measure of the skew
LinearTrend	Slope of a linear fit to the light curve
MaxSlope	Maximum absolute magnitude slope between two consecutive observations
Meanvariance	Ratio of the standard deviation to the mean magnitude
PairSlopeTrend	Considering the last 30 (time sorted) measurements of source magnitude, the fraction of increasing first differences minus the fraction of decreasing first differences
PeriodLS	Period corresponding to frequency of maximum power in the Lomb–Scargle periodogram
PercentAmplitude	Largest percentage difference between either the max or min magnitude and the median
Psi_eta	η^e index calculated from the phase-folded light curve
SmallKurtosis	Small sample kurtosis of the magnitudes

Table 3. Supplementary data from *Gaia* EDR3 incorporated as data set features (see Section 2.3).

Feature	Description
ra, dec, ra_error, dec_error	Right ascension, declination, and associated standard errors
l, b	Galactic longitude and Galactic latitude
ecl_lon, 'ecl_lat	Ecliptic longitude and ecliptic latitude
bp_rp, bp_g, g_rp	BP − RP, BP − *G*, and *G* − RP colours
*phot_***X***_mean_flux*	Mean flux in the *G*, integrated BP, or integrated RP bands – corresponding to **X** = *g, bp, or rp*, respectively
*phot_***X***_mean_flux_error*	Error on the mean flux in the **X** band
*phot_***X***_mean_flux_over_error*	Mean flux in the **X** band divided by its error
*phot_***X***_mean_mag*	Mean magnitude in the *G*, integrated BP, or integrated RP bands – corresponding to **X** = *g, bp, or rp*, respectively
pseudocolour, pseudocolour_error	The astrometrically estimated effective wavenumber of the photon flux distribution in the astrometric *G* band, measured in μ^{-1}m, and standard error of pseudo-colour
parallax, parallax_error	*Gaia* parallax in milliarcseconds (mas) and standard error
parallax_over_error	Parallax divided by its standard error
pm, pmra, pmdec	Total proper motion, and proper motion in the right ascension and declination directions (mas yr^{-1})
pmra_error, pmdec_error	Standard error of the proper motion in right ascension and declination directions (mas yr^{-1})
ruwe	Renormalized unit weight error: expected to be around 1.0 for sources where the single-star model provides a good fit to the astrometric observations. A value significantly greater than 1.0 (say, >1.4) could indicate that the source is non-single or otherwise problematic for the astrometric solution

assigned to examples such that incorrect predictions of the tree are given higher weights than those correctly predicted. This iterative process of modifying weights before training means that difficult-to-predict examples are given more influence. XGBoost is another example of sequentially combining decision trees. Where ADB uses weights to improve performance, XGBoost aims to reduce some error function that describes the classification performance of successive trees (Chen & Guestrin 2016).

SVMs (Cortes & Vapnik 1995) find the ideal hyperplane that best distinguishes between two classes in feature space. For non-linear class separation, SVM uses the *kernel trick*, transforming lower dimension input feature space into a higher dimensional space allowing for linear separation. KNN (Zhang 2016) stores the position vectors of training set class examples in feature space. Class predictions on new examples are made by assigning the mode of the classes of the KNN from the training set to the new example. ANN (LeCun, Bengio & Hinton 2015) consists of layers of interconnected nodes (or neurons) – an input layer, consisting of feature values; an output layer, which delivers the predictions (e.g. class probabilities); and, in between, one or more hidden layers of neurons, which sequentially transform the feature values into the predictions by applying typically non-linear functions to linear combinations of prior inputs. The algorithm learns through a process of loss minimization whereby the model parameters are adjusted to reach convergence to loss minimum (known as back propagation).

Table 4. The hyperparameters explored for each ML algorithm.

RF hyperparameters	Description
n_estimators	Number of decision trees
max_features	Maximum number of features provided to each tree
max_depth	Maximum number of binary split levels in each tree
ADB hyperparameters	
n_estimators	Same as for RF
learning_rate	Weight assigned to each classifier at each boosting iteration. This determines the impact of each tree on the final outcome
max_depth	Same as for RF
XGBoost hyperparameters	
n_estimators	Same as for RF
min_child_weight	Minimum sum of weights of all observations in a child node
gamma	Nodes are split only when there is a reduction in the error defined by a loss function. Gamma specifies the minimum loss reduction required to make a split
Subsample	Fraction of examples to be randomly sampled for each tree
colsample_bytree	Similar to *max_features* in RF
max_depth	Same as for RF
SVMs hyperparameters	
Kernel	See text: 'radial basis function (RBF)'
Kernel coefficient (γ)	Defines how far the influence of a single training example reaches, where the values can be seen as the inverse of the radius of influence
Error penalty (C)	Controls the cost of misclassification on the training data. Small C = soft margin, large C = hard margin
KNNs hyperparameters	
n_neighbours	Number of nearest neighbours to use
MLP hyperparameters	
learning_rate	Controls how much to change the model in response to the error each time the model weights are updated
Number of hidden layers	Number of hidden layers
Number of neurons	Number of units (neurons) within a given hidden layer
Activation function	Converts the output of a neuron into a form that serves as input for the next. Used to introduce non-linearity to a network

3.2 Fine-tuning

Controlling how the algorithms learn from the data set to generate predictive models is done by tuning their hyperparameters. They are algorithm settings that are set prior to the learning process. The aim with hyperparameter tuning is to improve model performance whilst the risk of overfitting (i.e. learning the noise in the data) is reduced. The hyperparameters explored for each of the algorithms tested are as given in Table 4.

3.3 Classification tasks

The classes assigned by the *Gaia* team are not mutually exclusive. For example, quasi-stellar objects (QSOs) are extremely luminous AGN. From the *Gaia* assigned classes, many variations of class grouping could be put forward for ML classification algorithms to distinguish. Two such groupings are defined by the following classification tasks, listed as transient class followed by the number of data set samples in brackets.

(i) *Binary classification* – CV (613) or not CV (4084).
(ii) *Four-class classification* – this comprises of the most populous transient types in the data set: AGN (which includes QSOs and BL Lac) as a single class (929), CVs (613), all different SNe (2713), and YSOs (184).

The tasks are assigned to the ML algorithms and their performance is evaluated. The classification tasks were first performed with both the light-curve extracted features and supplementary features. However, between 58 per cent and 90 per cent of data are missing for supplementary features. This was either due to unsuccessful cross-matching of targets with EDR3 – cross-matching was unsuccessful for 90 per cent of SNe, 23 per cent of CVs, and <1 per cent of AGN and YSOs, respectively – or certain metadata not being available where cross-matching was successful. For example, parallax measurements may not be available if the target is too faint or distant for an accurate measurement. Therefore we felt it necessary to also perform classification tasks with light-curve extracted features alone. These implementations can then be compared with other works where classification has been performed using light-curve-derived features alone.

3.4 Data pre-processing

Prior to ingestion into ML algorithms, the associated data sets require some level of preparation. Examples within each task specific data set contain missing data for several features. The strategy employed here is to replace missing values with the mean value of the feature column (mean imputation; Khan, Khan & Singh 2018). Feature scaling is employed for all except the ensemble learning algorithms (i.e. RF, ADB, and XGBoost) so that features with a larger range of values do not impart more influence on the model during training and for faster convergence to error minimum for gradient descent algorithms. We standardized the data to achieve zero mean and unit variance (or equivalently, standard deviation; Muhammad Ali & Faraj 2014).

3.5 Train–test split

Training and evaluation of a ML model require a separate training and test set. The algorithms are trained on the training set to generate

a model to be evaluated on the test set. The task specific data sets are split 50/50 into a training set and test set in a stratified manner – the same proportion of each class is represented in each of the training and testing sets. The split is performed before the pre-processing (imputation and feature scaling) stages to avoid information from the test set being present within the training set (data leakage) and yielding extremely biased results on model performance.

3.6 Optimal hyperparameter search

Model performance depends significantly on the selection of hyperparameters. Testing all combinations manually for the optimal combination is unfeasible. Therefore the GridSearchCV and RandomisedSearchCV functions from scikit-learn's (Pedregosa et al. 2011) model_selection PYTHON package were used to loop through combinations of predefined hyperparameters to identify an optimal set of hyperparameters for a given model. The cross-validation method is used to evaluate the performance of each combination. Cross-validation involves randomly dividing a set of observations into k groups, or folds, of approximately equal size. For each unique fold, the algorithm is trained and a model built on k-1 folds, this model is then evaluated on the remaining fold (validation set) with a chosen performance metric. This is repeated until each of the k-folds has served as a validation set. The average of the k recorded accuracies (or other chosen metric) is the cross-validation score serving as the model performance metric. For this work a 10-fold cross-validation was implemented. The training set was entered into these functions and the best combination of hyperparameters for a given algorithm was found. The model with this combination was then evaluated on the test set.

3.7 Classifier performance

The number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) serves as the basis for performance metrics used to assess model performance. These quantities first require the definition of the positive class, this is the class of interest (CV in this case); and the negative class, non-CVs in the binary classification task, and all other classes in the multiclass case.

3.7.1 Confusion matrix

It is common to present the redTP, TN, FP, and FN quantities in an $N \times N$ table known as a *confusion matrix* (where N represents the number of classes). This construct allows us to easily identify the number of TPs, TNs, FPs, and FNs. These values may then be used to evaluate the performance of our binary and multiclass ML models using the class specific precision, recall, and F1-score; and the overall model accuracy and balanced accuracy.

3.7.2 Precision, recall, and F1-score

The precision is defined as the fraction of examples our model predicted as belonging to the positive class that do actually belong to this class: TP/(TP + FP). In other words, it tells us how much we can trust our models predictions of the positive class. The recall is the fraction of examples of the positive class that our model correctly predicted as belonging to this class: TP/(TP + FN). This metric assesses the model's ability to identify all members of the positive class. The F1-score is the harmonic mean of precision and recall for our positive class, and is useful in finding the best trade-off between these quantities. The highest possible value of the F1-score is 1 (100 per cent), indicating perfect precision and recall, the lowest possible value (0) relates to a score of 0 for either precision or recall.

3.7.3 Accuracy and balanced accuracy

The accuracy is the fraction of all examples whose class was correctly predicted by the model. For binary classification this is (TP + TN)/(TP + TN + FP + FN), for a multiclass situation we sum the number of true positives for each class and divide by the total number of examples. The accuracy returns an overall measure of the model's predictive capability. Should we only be concerned with assigning the most number of examples to their correct class, accuracy is a good metric. However, under this metric, strong classification errors for classes with few examples to their name will be hidden. Therefore, should we be concerned with finding a model that has a strong classification performance across all classes, we may use 'balanced accuracy' that can account for this class imbalance. This is the arithmetic mean of the recalls for each class.

3.7.4 Area under the curve of the receiver operating characteristic

The receiver operating characteristic (ROC) curve allows us to see the trade-off between sample purity and completeness plotted as the true positive rate (TPR; also called recall) as a function of the false positive rate (FPR; the fraction of examples incorrectly predicted as belonging to the positive class: FP/(TN + FP)) for all values of a threshold probability above which a positive classification is made. More specifically, for each example, algorithms return a probability of belonging to a certain class. The probability is thresholded such that examples with probabilities equal or greater than the threshold are mapped to the positive class and remaining examples are mapped to the negative class. Therefore the ROC curve is an evaluation of TPR and FPR as we continuously vary this probability threshold. This can be used to determine the appropriate threshold for a given study such that we can adjust for our desired level of purity and completeness depending on our science case. The goal of classification is to maximize the TPR while minimizing the FPR. For the binary classification tasks, the area under the curve (AUC) of the ROC can be used to assess model performance. A value of 1 for the AUC indicates a perfect model, capable of correctly assigning the correct class prediction to all examples, when AUC = 0.5, the model is no better than a random guess, while AUC = 0 corresponds to incorrectly predicting classes of all examples.

4 RESULTS

Tables 5 and 6 show the model evaluation scores for the binary and four-class classification tasks. The scores for models trained with both light-curve extracted and supplementary features (full feature models) are shown without brackets, while the scores for models trained with light-curve extracted features alone (light-curve only models) are within brackets. The scores shown are the accuracy, balanced accuracy, and with respect to the CV subclass, the precision, recall, and F1-score. The choice of best performing model is based on the F1-score for the CV class. This metric was chosen as it considers both the need to minimize FPs, which is important for the efficient use of telescope time for target follow-up, and a requirement to minimize FNs.

Table 5. Binary task classification scores for ML models as measured on the test set. Scores without brackets relate to models using both light curve and supplementary features, while those in brackets are for models that used only light-curve extracted features. Random forest (RF) was implemented with 100, 250, 750, and 1000 trees denoted by RF then the number of trees; other abbreviation are ADA – AdaBoost, MLP – multilayer perceptron, KNN – k-nearest neighbour, and SVM – support vector machine.

Model	Accuracy	Balanced accuracy	CV precision	CV recall	CV F1-score
RF100	0.955	0.870	0.88	0.76	0.81
	(0.938)	(0.824)	(0.82)	(0.67)	(0.74)
RF250	0.955	0.870	0.89	0.75	0.81
	(0.939)	(0.829)	(0.82)	(0.68)	(0.74)
RF500	0.955	0.868	0.89	0.85	0.81
	(0.937)	(0.827)	(0.81)	(0.68)	(0.74)
RF750	0.955	0.867	0.89	0.75	0.81
	(0.937)	(0.826)	(0.81)	(0.67)	(0.74)
RF1000	0.956	0.870	0.90	0.75	0.82
	(0.938)	(0.826)	(0.82)	(0.67)	(0.74)
ADA	0.959	0.874	0.91	0.76	0.83
	(0.932)	(0.840)	(0.75)	(0.72)	(0.73)
XGBoost	0.962	0.883	0.92	0.78	0.84
	(0.943)	(0.823)	(0.86)	(0.67)	(0.76)
MLP	0.932	0.824	0.78	0.68	0.72
	(0.932)	(0.822)	(0.78)	(0.67)	(0.72)
KNN	0.909	0.812	0.65	0.68	0.66
	(0.900)	(0.812)	(0.60)	(0.69)	(0.64)
SVM	0.817	0.787	0.39	0.75	0.52
	(0.871)	(0.802)	(0.51)	(0.71)	(0.59)

Table 6. Four-class classification scores. Score with and without brackets, and abbreviations are as described in Table 5.

Model	Accuracy	Balanced accuracy	CV precision	CV recall	CV F1-score
RF100	0.964	0.941	0.92	0.85	0.88
	(0.922)	(0.835)	(0.80)	(0.79)	(0.80)
RF250	0.964	0.936	0.92	0.85	0.88
	(0.924)	(0.835)	(0.81)	(0.79)	(0.80)
RF500	0.965	0.941	0.92	0.85	0.88
	(0.923)	(0.830)	(0.80)	(0.80)	(0.80)
RF750	0.965	0.942	0.92	0.86	0.89
	(0.923)	(0.833)	(0.80)	(0.80)	(0.80)
RF1000	0.965	0.942	0.92	0.86	0.89
	(0.923)	(0.833)	(0.80)	(0.80)	(0.80)
ADA	0.959	0.925	0.90	0.82	0.86
	(0.897)	(0.798)	(0.75)	(0.75)	(0.75)
XGBoost	0.962	0.928	0.91	0.84	0.87
	(0.922)	(0.820)	(0.83)	(0.77)	(0.80)
MLP	0.926	0.873	0.80	0.76	0.78
	(0.910)	(0.793)	(0.73)	(0.71)	(0.76)
KNN	0.895	0.795	0.90	0.48	0.63
	(0.898)	(0.730)	(0.86)	(0.67)	(0.75)
SVM	0.891	0.798	0.62	0.70	0.66
	(0.874)	(0.758)	(0.76)	(0.61)	(0.68)

4.1 Binary classification

4.1.1 Full feature model

The best performing binary task full feature model was XGBoost, trained with 150 decision trees at a learning rate of 0.1 and maximum tree depth of 6. The model outperformed each of the others with an F1-score of 84 per cent, the AdaBoost and RF implementations follow closely behind (81–83 per cent). There is though little difference between the top two models; use of the McNemar's test to compare the XGBoost and AdaBoost models show both classifiers make errors in much the same proportion (for $\alpha = 0.05$; $p = 0.175$). The confusion matrix (left-hand panel of Fig. 1) indicates 69 of the 307 CVs in the test set were misclassified by the XGBoost model, while of the 258 examples predicted as CVs only 20 were not. The corresponding ROC curve is plotted in the left-hand panel of Fig. 2, with AUC score of 0.975. The importance of each feature for a given model can be given by the feature importance scores. The 20 features with the largest effect on the model's predictive accuracy are plotted in the left-hand panel of Fig. 3. The number of observations greater than 2 mag brighter than the median of a rolling window has by far the greatest influence in discriminating between the classes.

4.1.2 Light curve only model

The best performing binary task light curve only model was XGBoost (CV F1-score of 76 per cent). The implementation was performed with 150 decision trees at a learning rate of 0.2 and maximum tree depth of 6. The RF models follow closely behind (CV F1-score of 74 per cent); use of McNemar's test again showing XGBoost makes errors in the same proportions ($0.766 \leq p \leq 0.88$). The CV F1-score performance for this XGBoost model drops compared to the full feature model by 8 percentage points due to an increase in the number of FN from 69 to 93 and an increase in the number of FP to 36 from 20. Out of the 307 test set CVs, 214 were correctly identified (see right-hand panel of Fig. 1). The model AUC score also drops from 0.975 to 0.9622 (see right-hand panel of Fig. 2). The number of observations greater than 2 mag brighter than the median of a rolling window remains the feature that has by far greatest influence in discriminating between the classes (right-hand panel of Fig. 3).

4.2 Four-class classification

4.2.1 Full feature model

A 750-tree RF model performs equally well or better than its competitors in each of the performance metrics evaluated for this four-class full feature task. The F1-score for CV classification stands at 89 per cent though the remaining ensemble learning models follow closely behind. The model was trained such that only 25 per cent of features (selected at random) could be used within each tree, with a maximum tree depth of 25. The confusion matrix (left-hand panel of Fig. 4) displays the strong performance in distinguishing CVs from other classes. Those CVs that were misclassified were mostly predicted to be of the SNe class (39/44). The left-hand panel of Fig. 5 presents histograms of the probabilities of class assignment for this model. The vast majority of test set examples, 274 out of the 285 predicted CVs, were predicted as such with probabilities greater than 50 per cent. 79 of the 285 were predicted as CVs with a probability of 95 per cent or above. All but three examples predicted as YSOs are classified with probability of 50 per cent or higher.

According to the feature importances (left-hand panel of Fig. 6) the temporal baseline of observations (*detected_time_diff*) has the greatest influence in discriminating between classes. In addition to *Gaia*'s observing strategy and their prevalence in the data set, this can be partially explained by the properties of the majority class, SN – they are too distant for their progenitors to be observable by *Gaia*, after several months they become too faint to be observable above the light from their host galaxy. Of the supplementary features, parallax

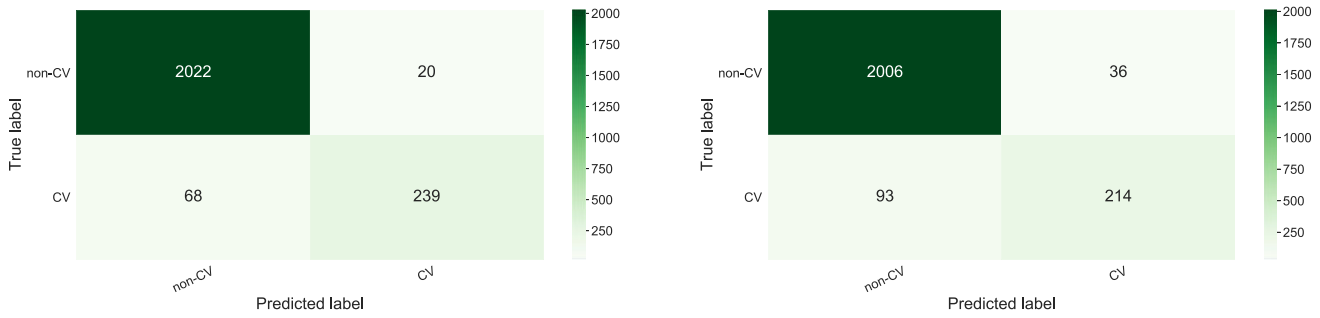

Figure 1. Confusion matrices (CMs) for the best performing binary task full feature (left) and light curve only (right) models. In each case this was an XGBoost model – achieved the highest F1-score. The CMs show the numbers corresponding to precision, recall, and accuracy scores in Table 5. There are over six and a half times more non-CVs in our test set than CVs, raising the overall accuracy score, the balanced accuracy score is more able to account for this class imbalance.

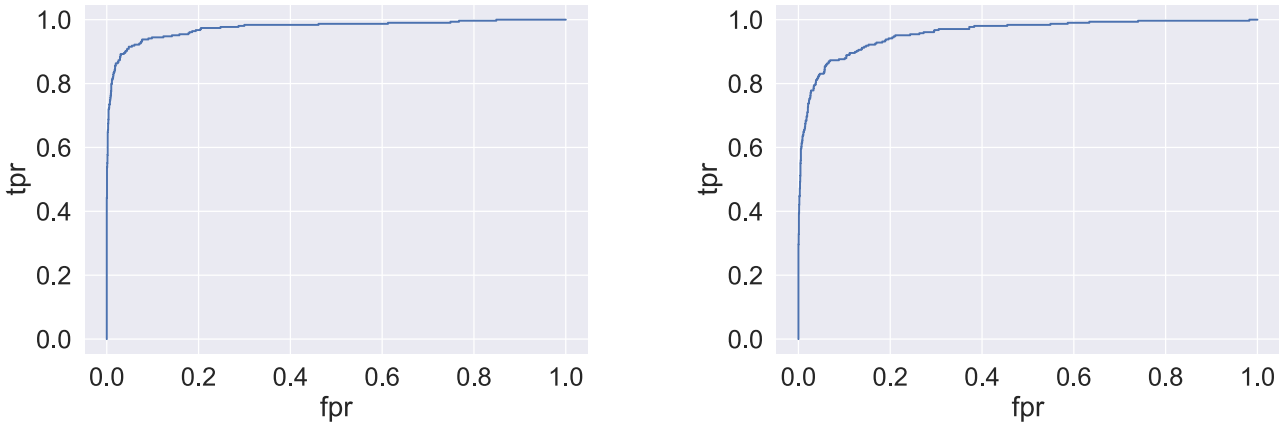

Figure 2. ROC curves for the full feature and light curve only binary task models achieving the highest CV F1-scores. On the left is the curve for the full feature model, while on the right is that for the light curve only model. The full feature model area under the curve is 0.975, for the light curve only model this is 0.9622, indicate a strong performance in each case.

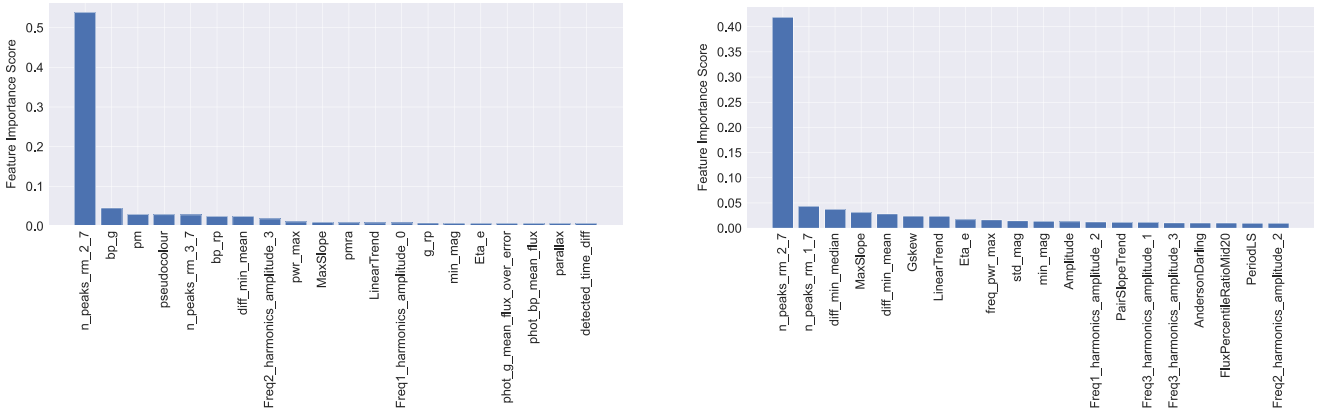

Figure 3. Feature importance scores for the 20 most influential features within the best performing full feature and light curve only binary task models. Feature importance refers to a class of techniques for assigning scores to input features to a predictive model, in this case XGBoost, that indicates the relative importance of each feature when making a prediction. The most important feature for each of the full feature (left) and light curve only (right) models is *n_peaks_rm_2_7* – number of instances of data points at least 2 mag brighter than the median of a rolling window of seven epochs. Feature definitions are contained in Tables 1–3.

and proper motion are expected to provide the greatest ability in class distinction, with the ability to distinguish extragalactic sources from those nearby. They both appear high in feature importances, as do the right ascension and declination error features. These errors are noticeably higher for SNe (∼12.8 mas) than for remaining classes (∼0.08–0.17 mas) attributed to the ability to measure these properties being affected by crowding (including contamination of light from the host galaxy).

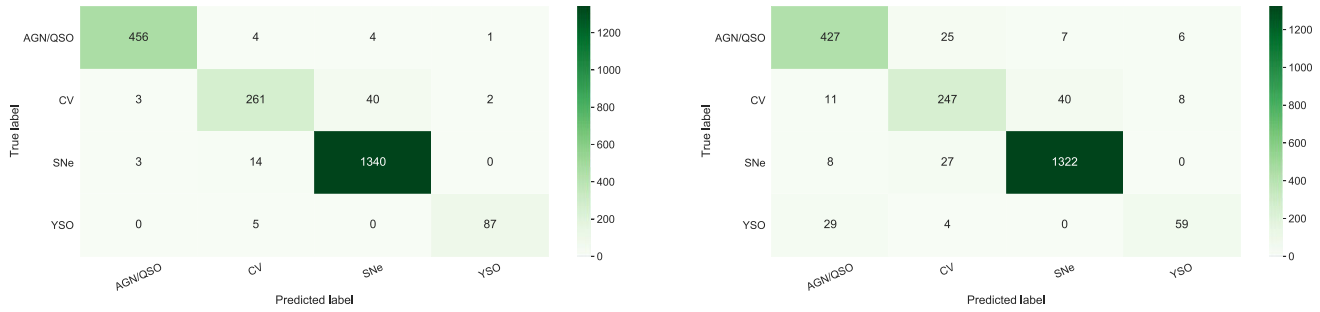

Figure 4. Confusion matrices for the best performing full feature and light curve only models in the four-class classification task. On the left is the 750-tree RF model trained with full complement of features. 262 of the 306 CVs in the test set were successfully classified (true positives), the majority of those misclassified, 39 of 44, were predicted to be SNe. On the right is the 1000-tree RF model trained with light curve derived features only. Less true positives (247) compared to the full feature model. Also an increase in the number of false positives from 23 to 56, of which the majority were AGN and SNe.

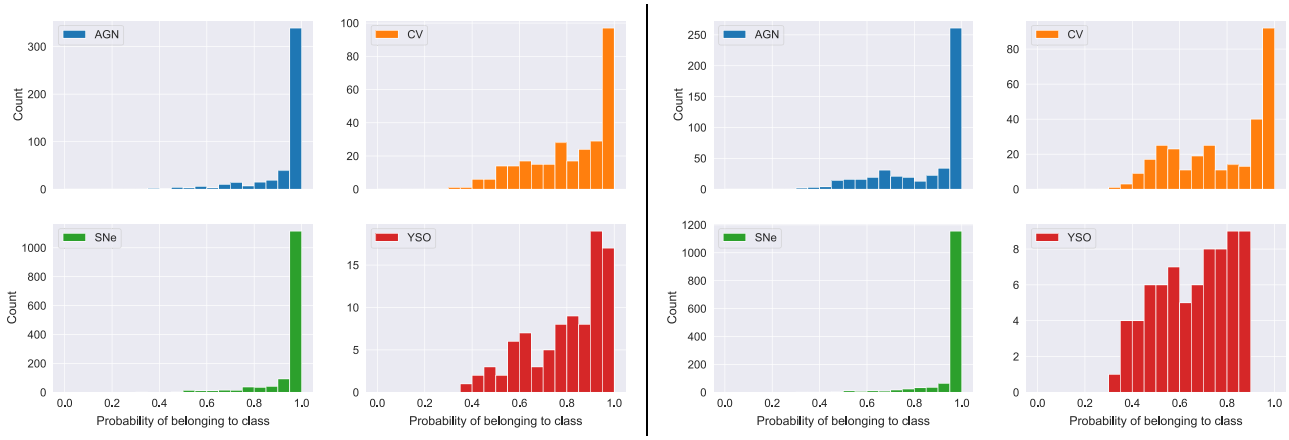

Figure 5. The number of test set examples predicted as CV (orange), AGN (blue), SNe (green), and YSO (red) separated in bins of probability of class association calculated for the full feature and light curve only four-class models with the highest F1-scores. Each tree in the RF model predicts class probabilities for each example – these are the fraction of samples of the same class in the associated leaf evaluated during training. These probabilities are averaged for the forest prediction. Class probabilities for the full feature RF model (left) show nearly all examples are assigned classes with greater than 50 per cent probability, the majority which are in the 95–100 per cent bins. For the light curve only feature four-class model (right), we can say likewise, however the YSO class assignment probabilities are more uncertain.

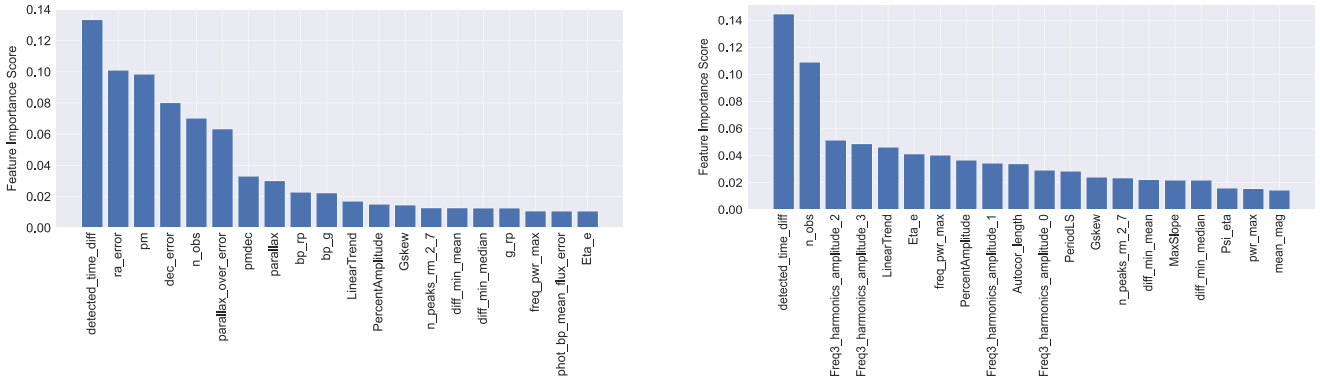

Figure 6. The top 20 features based on feature importance scores for the four-class full feature and light curve only models with the highest F1-scores. The full feature model best performing feature (left) was the time between the first and last observation of the target, followed by the error in the right ascension, proper motion, and the error in the declination. The same best performing feature is present for the light curve only model (right). Feature definitions are contained in Tables 1–3.

4.2.2 Light curve only model

A 1000-tree RF model performed the best, achieving the highest CV F1-score (81 per cent) for the four-class light curve feature only task, though the remaining ensemble learning models follow closely behind. The model was implemented with a maximum tree depth of 30 and 75 per cent of randomly selected features available for each tree. While 247 of the 306 CVs have been correctly classified (right-hand panel of Fig. 4), the contamination

of other classes into those targets predicted as CV increases from 8 to 18 per cent compared to the full feature model. Like the full feature model, misclassified CVs are mostly assigned the SNe label. The histograms of class assignment probability (right-hand panel of Fig. 5) show the majority of CVs are predicted as such with greater than 50 per cent probability, though more examples are now present in the tail of the distribution. According to the feature importances (right-hand panel of Fig. 6) the temporal baseline of observations (*detected_time_diff*) also has the greatest influence in class distinction.

5 DISCUSSION

5.1 Semi-regular, short-duration outbursts

The light-curve feature that logs the number of epochs that are at least 2 mag brighter than the median of a rolling window of seven epochs, *n_peak_rm_2_7*, outperforms all others in feature importances for the best performing full feature and light curve only binary models. The semi-regular, short-duration outbursts of DNe are effectively picked out using this feature as found during its development. Such characteristics are more likely to be identified within *Gaia*'s transient alerts pipeline than the less frequent alert triggering features of other CV subtypes, so the high ranking of the feature may be expected. Indeed, a coordinate cross-match with 'The Catalogue and Atlas of Cataclysmic Variables'[5] (Downes & Shara 1993; Downes, Webbink & Shara 1997) reveals 77 per cent of successfully cross-matched data set CVs are of listed as being DNe. Exploration of the true positives for each of the best performing binary models reveals the majority display the expected DNe morphology (78 per cent and 73 per cent for the full feature and light curve only models, respectively).

5.2 Limited epoch photometry

A significant fraction of our data set is constructed from target light curves with few epochs of observation, 36 per cent of targets contain five or fewer data points in their light curves. This is due to the combination of *Gaia*'s sampling frequency and systems too faint to be observed by *Gaia* until a brightening event propels them into visibility. Transient phenomena more likely to display this trait will be those exhibiting a rapid and large amplitude brightening, for example SNe and the CV subclasses of classical and DNe. Considering SNe comprise the majority (58 per cent) of our data set, this may provide an explanation for the strong performance of *detected_time_diff* (temporal baseline of observations) in class distinction (see Fig. 6). It may also explain the difficulty that the best performing four-class models have in distinguishing CVs from SNe. Of the CVs misclassified by the best performing full feature four-class model, 87 per cent (39 of 45) are predicted to belong to the SN class, while for the corresponding light curve only model, 68 per cent (40 of 59) are predicted as SNe. Similarly, the majority of misclassified SNe in each of those models are predicted to be of the CV class. Inspection of the CVs misclassified as SNe reveals the majority possess light-curve morphologies that are present for the SNe samples – those with few data points (2–10 observations) and those exhibiting an approximately exponential decline with no pre-explosion data.

[5] https://archive.stsci.edu/prepds/cvcat/index.html

5.3 Metadata and high imputation

A McNemar's test suggests the use of metadata has an impact on model performance when comparing the full feature and light curve only XGBoost models ($p = 10^{-7}$). However, the small difference in classification accuracy between these two binary models (1.9 per cent) indicates that the addition of survey metadata provides minimal benefit in distinguishing CVs from non-CVs. This is also shown by the small difference in the AUC (1.3 per cent) between these models, with both performing strongly by this measure (0.975 and 0.962 for the full feature and light curve only models, respectively). The feature importances for both binary models further illustrate this point – the influence of supplementary features in class distinction is dwarfed by the light-curve-derived *n_peaks_rm_2_7* feature. Either the metadata is unimportant or mean imputation has diluted the influence this data has on class distinction. The latter seems more likely when presented with pair plots of Fig. 7 that show transient classes in metadata feature space. This plot is of particular use in interpreting the performance of algorithms that rely on class separation within feature space (e.g. KNNs and SVMs). Evident is the distinction between YSOs from CVs, SNe, and AGN in colour space (bp − rp, bp − g, g − rp); and CVs and YSOs from SNe and AGN when proper motion is considered.

Use of mean imputation has its drawbacks, it ignores relationships between features, the correlation for example, and reduces the variance of the variable, thereby introducing bias to our model. Furthermore, the strategy may not be suitable for several supplementary features. For example, the parallax may not be measurable because the object is too far away (too small to measure); and a missing value for proper motion can either be due to the object having no proper motion to measure or be due to it being too distant to be measured. A more appropriate strategy could be to replace these with a value of zero – a more accurate quantity for the parallax and proper motion of the most distant sources – though this does not account for the unavailability of these features due to an unsuccessful cross-match with EDR3. While alternative methods of handling missing data could be employed (such as those summarized in Soley-Bori 2013), a large amount of data is missing for the supplementary features (58–90 per cent), this can limit the effectiveness of any such strategy (Jäger, Allhorn & Bießmann 2021). Fig. 7 shows how photometric colour information can be an important property for class distinction, this is readily available in multiband surveys such as ZTF and can be used to help alleviate the issue.

5.4 Comparison with other work

The results of our investigation compare favourably with similar classification attempts where CVs are included as class. Neira et al. (2020) experiments with CRTS light curves in their eight-class classification model yielded an F1-score of 75 per cent for the CV class, while this work exceeds this in both the binary (76 per cent) and four-class (80 per cent) tasks where only light-curve features are used. Sánchez-Sáez et al. (2021) evaluated three different algorithms in their tiered classification attempts to distinguish between CVs, SNe subclasses, AGN, YSOs, and variable star subclasses from a data set constructed from ZTF light curves and colours from AllWISE. Their CV recall scores for their implementation of the balanced RF (Chen, Liaw & Breiman 2004), XGBoost, and MLP classifiers are 68 per cent, 72 per cent, and 61 per cent, respectively. This compares with 67 per cent and 80 per cent for our light curve only best performing binary and four-class models, respectively. These comparisons do not however

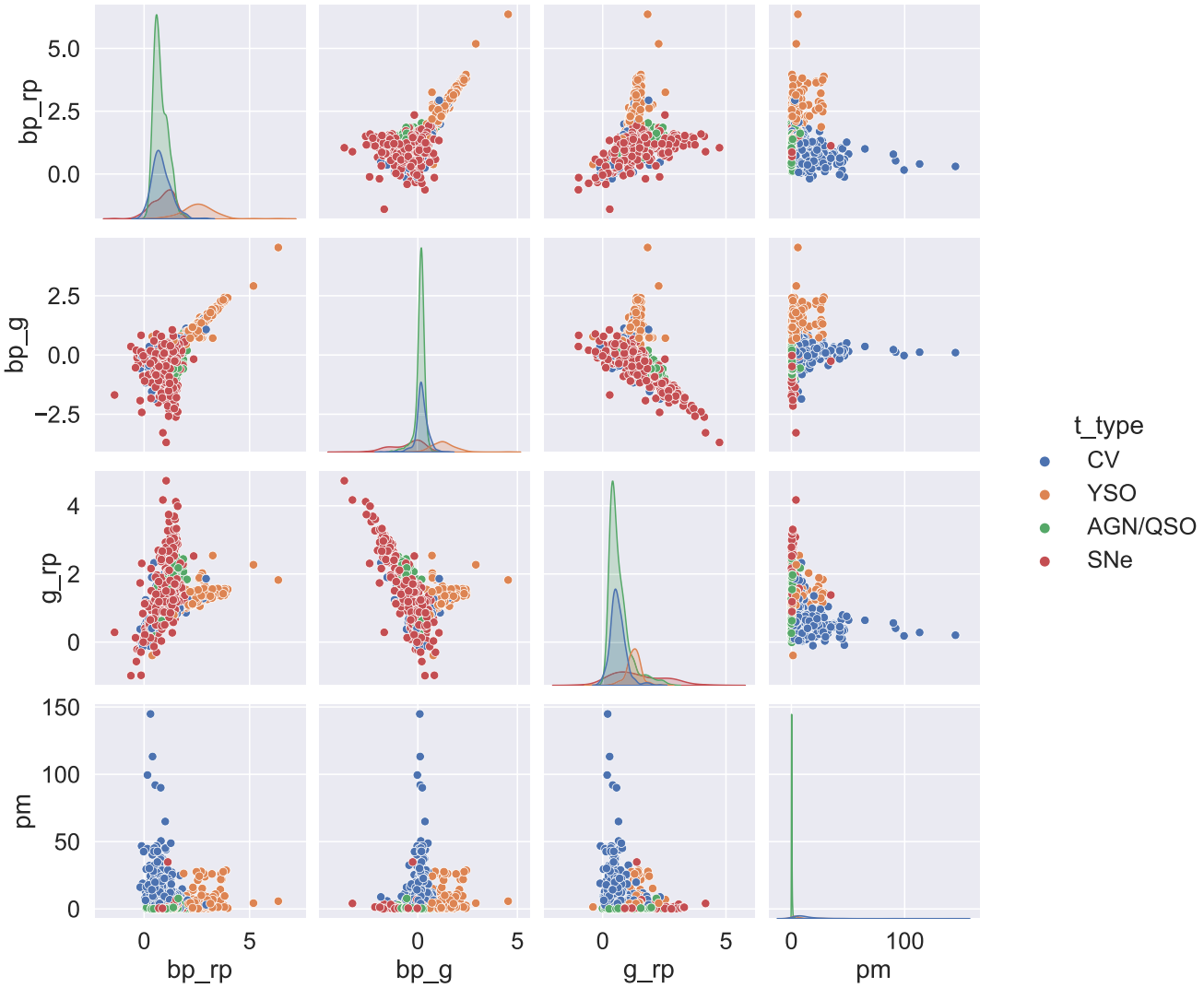

Figure 7. The pairplot allows us to see both the distribution of the single variables (plots shown diagonally from top left to bottom right) and relationships between two variables (off diagonal plots). This is shown for the bp − rp, bp − *g*, and *g* − rp colours, and proper motion. YSOs are redder in colour compared to SNe, CVs, and AGN, observed in both the single variable distribution and relationship plots, thus allowing for a significant level of class separation. Introduction of proper motion allows for separation of the more distant (extragalactic) SNe and AGN from the closer (Galactic) CV and YSO population.

take into account differences in the instruments used to collect the data, which translates to the nature of photometric data (e.g. observing cadence, waveband). Furthermore comparisons do not consider differences in transient classes to classify and ML methods employed.

5.5 *Gaia* unknowns

5.5.1 Model predictions on unknown sample

The model that produced the highest CV F1-score overall – full feature four-class model (RF with 750 trees) – is used to make class predictions of targets labelled as 'unknown' (unclassified) within the *Gaia* alerts stream. As of 2021 December, 13 241 targets were of 'unknown' class. Of these, the model predicted 2833 (21 per cent) to be of the CV class, 1928, 6611, and 1869 were classified as AGN/QSO, SNe, and YSOs, respectively . As mentioned in Section 2, the unknown sample will contain several minority classes (e.g. microlensing and tidal disruption events) not included in the test set used to evaluate model performance. We aim to assess the impact this has on our model's ability to generalize to the unknown sample, and new transient alerts in general. This will require spectroscopic observations for a statistically significant fraction of our 2833 predicted CVs to identify their true transient classification.

5.5.2 Spectroscopic follow-up

We are therefore undertaking a pilot study to assess the performance of our model and the methods used by obtaining spectroscopic observations to classify those targets that can be observed with the Spectrograph for the Rapid Acquisition of Transients (SPRAT) low-resolution spectrograph (Piascik et al. 2014) mounted on the Liverpool Telescope (LT; Steele et al. 2004). These spectra cover

Table 7. Classifications based on the Liverpool Telescope (LT)/Spectrograph for the Rapid Acquisition of Transients (SPRAT) spectroscopy of several targets labelled as 'unknown' (without a transient class assignment) within *Gaia* Science Alerts (GSA) and predicted as CV by the RF750 model.

Target	Classification	Comment
*Gaia*16cfn	CV (DN)	Clear Balmer and He I emission. He I λ4922 blended with Fe II λ4924. Characteristic of DNe subtype
*Gaia*17ccv	CV (decline from DN outburst)	CV on decline from outburst, faint Hα and Hβ emission, He II λ4686 in emission. Double peaked lines – indicative of high inclination system
*Gaia*17dfn	CV	Balmer and He I λ4471 lines in emission
*Gaia*18auz	CV	Clear Balmer emission with several faint He I lines in emission
*Gaia*18dgt	CV (DN)	Broad Balmer emission with lines of He I. He I λ4922 blended with Fe II λ4924. Characteristic of DNe subtype. Double peaked emission, possible high inclination system
*Gaia*18dhv	CV	Balmer, He I and He II in emission
*Gaia*19bzn	CV	Clear Balmer emission; faint lines of He I and He II
*Gaia*19cln	CV	Clear Balmer emission; lines of He I and He II also present; He I λ4922 blended with Fe II λ4924
*Gaia*20air	CV	Clear Balmer emission; lines of He I and He II also present; He I λ4922 blended with Fe II λ4924
*Gaia*20bjd	CV	Clear Balmer emission; lines of He I also present; He I λ4922 blended with Fe II λ4924
*Gaia*20cpq	CV (DN)	Clear Balmer emission; lines of He I and He II also present; He I λ4922 blended with Fe II
*Gaia*21beh	CV	Outburst spectrum. Possible very faint Hα absorption, clear absorption in remaining Balmer lines and He I λ4471, He II λ4686 in emission (faint)
*Gaia*21cgv	CV	Balmer and He I emission lines, faint Fe II λ5169
*Gaia*21cul	CV	Clear Balmer and He I emission lines
*Gaia*21eyb	CV	Balmer, He I, He II, and Fe II emission lines, He I λ4922 blended with Fe II λ4924

Table 8. List of sources list as being of 'unknown' type by GSA that our four-class full feature RF model predicted as belonging to the CV class. The full table of 2833 sources is available as supplementary material online.

Gaia object name	Right ascension	Declination
*Gaia*17avy	343.52368	65.09345
*Gaia*18cdn	263.44916	− 30.53935
*Gaia*20cjd	223.95987	− 64.73914
*Gaia*20eno	271.03934	− 76.22711
*Gaia*21bnn	50.13837	6.71808
*Gaia*20fax	338.65765	58.42598
*Gaia*20efe	120.77764	− 17.40920
*Gaia*19dal	312.73430	31.95018
*Gaia*21bqw	120.07837	− 38.82770
*Gaia*21bby	291.18503	− 34.15307
...	...	...

a wavelength range of 4000–8000 Å with a resolution of 18 Å, corresponding to a resolving power $R = 350$ at the centre of this range. A limit on telescope time and the need for high-quality spectra require of an efficient observation strategy. Accordingly, observations are limited to targets with a median brightness no fainter than 18th mag. Furthermore, only those targets that rise highest in the sky – visible for longer at a lower airmass – are considered. Therefore we limit our sample to those with a declination corresponding to an altitude no lower than 50° when at transit altitude. These cuts leave a sample of 220 targets, 7.8 per cent of the total catalogue, representing a statistically significant fraction with which we can validate the performance of our model. We have spectroscopically classified 15 of this sample, all of which we can confirm are of the CV class. Details of these targets are given in Table 7, while the associated SPRAT spectra are shown in Fig. 8. Classification as a CV is based on the presence of Balmer and/or He I/He II lines. Where the signal-to-noise ratio of the spectrum permits, subtype classification is performed. Full details of the spectral features we used for classification are given in Szkody (1998) and Hou et al. (2020).

6 CONCLUSIONS AND FUTURE WORK

The advent of wide-field synoptic surveys has revolutionized time domain astronomy with their ability to detect millions of transient event per night. The use of ML is recognized as the best method of source classification for this deluge of transient sources. ML algorithms have been applied widely to data from several surveys including CRTS and ZTF photometry. In this work we applied ML techniques to the transient stream of GSA, a resource not fully explored with ML. Our focus lies in the identification of CV stars, a class of transients providing ideal laboratories for the study of accretion and binary evolution. Using features extracted from light curves of classified sources and associated metadata as input, we evaluated the use of RF, Adaboost, XGBoost, KNNs, SVMs, and an MLP in performing several tasks. These are the identification of CVs in the context of binary classification (CV or non-CV) and a four-class task (CV, AGN, SNe, and YSOs). Each of these tasks was performed with and without metadata (e.g. *Gaia* parallaxes and colour) during training. By comparison of the F1-score of all models across both tasks, the four-class RF model trained with both light curve and metadata-based features performed the best with an F1-score of 89 per cent when evaluated on the test set. We applied this model to the list of unclassified targets within GSA. The model predicted 2833 of these 'unknowns' to be of the CV class. We are now undertaking a spectroscopic observing campaign to spectroscopically classify a statistically significant sample of these targets to validate our models performance. So far we have been able to spectroscopically confirm 15 targets to be of the CV class.

The use of data beyond light curve features seems necessary in order to achieve classification performance close to F1 = 90 per cent. However, with light-curve features alone the performance of the model compared well with other works, despite more sparsely sampled light curves. The lessons learnt during this exploration of the GSA resource and the classified targets from our spectroscopic database of targets will be useful in the next phase of our research. This will be the application of ML to the multiband high-cadence light curves of the ZTF survey.

The next phase will be an opportunity to explore methods of handling class imbalance and missing data. Class imbalance, present

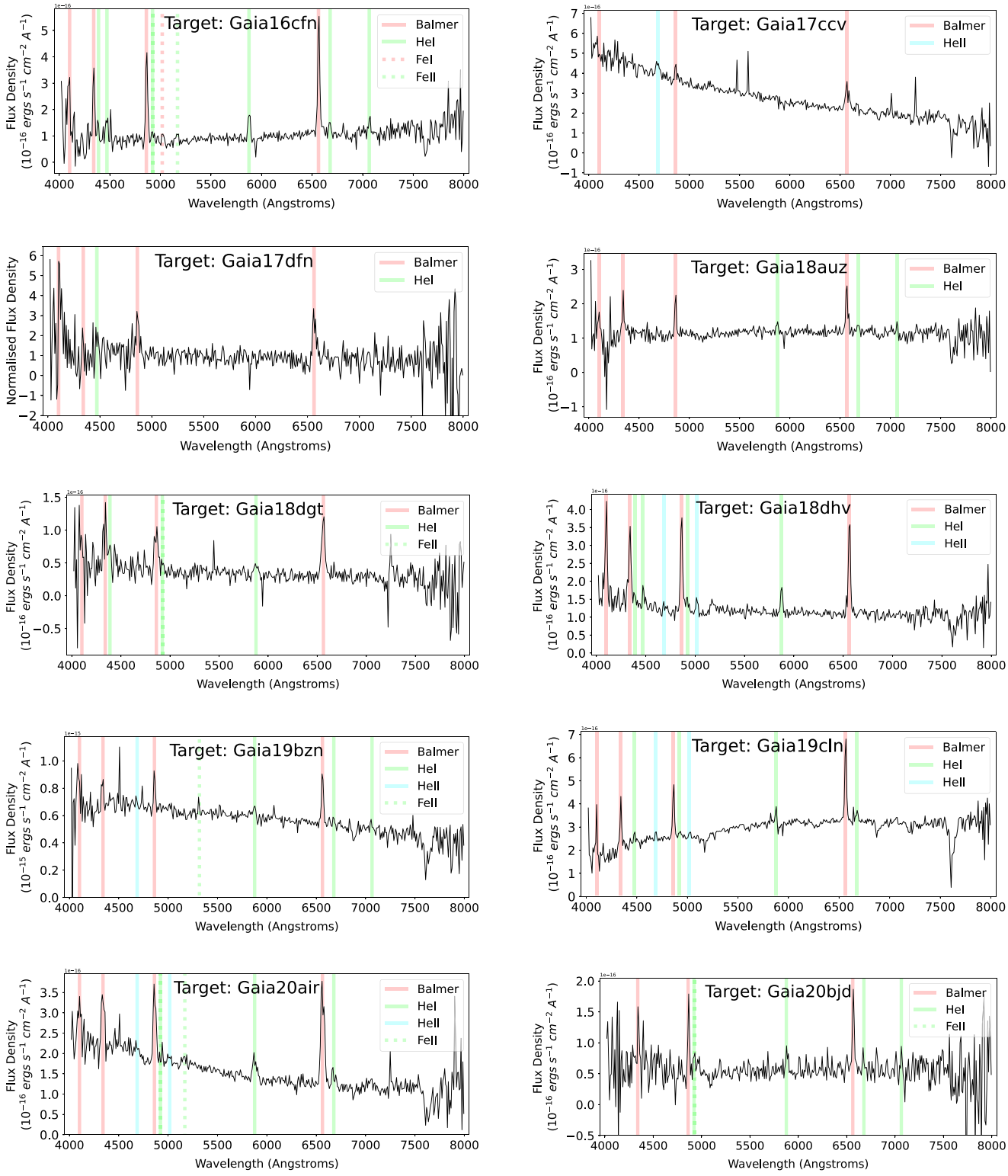

Figure 8. Spectrograph for the Rapid Acquisition of Transients (SPRAT) spectra of targets in Table 7. Spectral lines indicated in plots, labelled in the legend for each.

within our data set (see Section 2.1), tends to bias classifiers to recognize the oversampled class more than the undersampled class. Algorithm-specific solutions exist, for example, within RF one may grow each tree with the same number of targets per class by oversampling or undersampling using the bootstrap sampling process (Fernández et al. 2018). Data augmentation methods (e.g. Wen et al. 2020) to generate new examples based on existing examples will also be explored. Our use of mean imputation for handling missing data

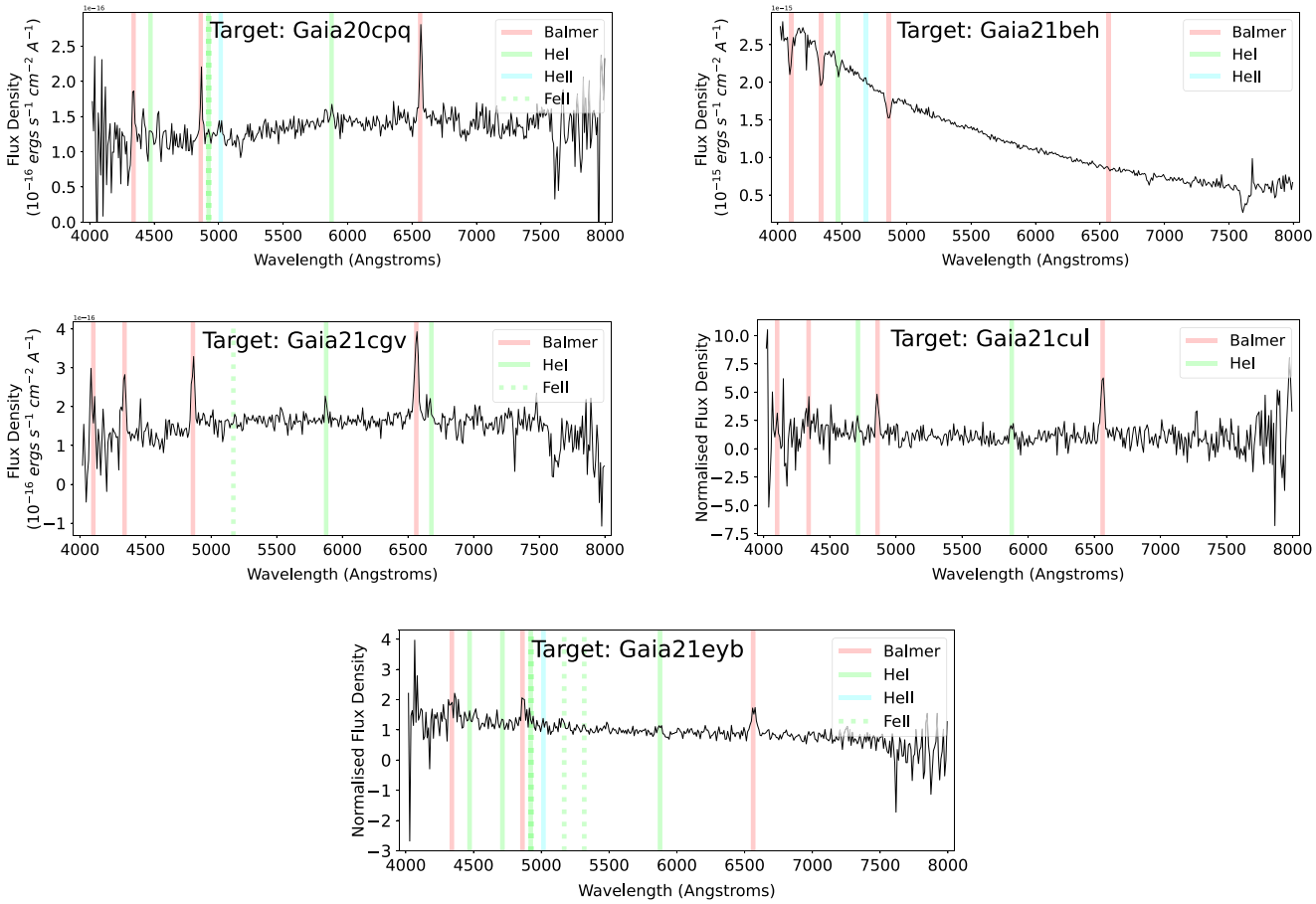

Figure 8. continued.

is simple and parameter free. Whilst this method can cause biases (see Section 5.3), we deemed the exploration of several imputation methods beyond the scope of this work, though it is something we will explore in work with ZTF data. The reliability of class labels will also be important for the next research phase and once LSST becomes operational. Whilst there is confidence in the methods employed in the labelling of examples used here (see Section 2.1), we acknowledge that labelling errors do occur. This can add noise to the data set, deteriorating classifier performance (Frenay & Verleysen 2014) and reducing the effectiveness of performance optimization techniques such as hyperparameter tuning.

The methods employed in this work are transferable to the data available from the ZTF survey. This data should provide the necessary information to identify subclasses within the CV population and pick out rare varieties that further our understanding of binary evolution. For example, the ∼2 d cadence provides the sampling necessary to recognize the defining characteristics of DNe subtypes, such as the superoutbursts of SU UMa systems (e.g. Szegedi et al. 2022) and the standstills of Z Cam systems (Simonsen et al. 2014); and identify characteristics present in outbursting AM CVns, such as the short-duration rebrightenings on the fading tail of a superoutburst (Kato & Kojiguchi 2021). Our ability to automatically distinguish between the different CV subtypes will depend upon several factors, one of which is the quality of features. Several features used in this work have so far shown their effectiveness at class distinction, others may become more significant once computed with the higher cadence data, while the development of features geared towards the identification of specific subtypes should provide further benefit. The prevalence of a given subtype within the data set is another factor that we expect to impact classifier performance. The sensitivity of a survey to certain CV subtypes results in the under-representation of novae, AM CVns and nova-likes compared to DNe due to the rarity of eruptions, faintness, and photometric stability, respectively. This is where the methods of handling class imbalance described in the previous paragraph will become invaluable. The methods used here and the lessons learnt will aid in our goal to separate the rare CV systems from those more common and hopefully lead to a greater understanding of binary evolution.

ACKNOWLEDGEMENTS

DM acknowledges a PhD studentship from Liverpool John Moores University (LJMU), Faculty of Engineering and Technology. MJD receives funding from UK Research and Innovation (UKRI) grant number ST/S505559/1. The Liverpool Telescope was funded by UKRI grants ST/S006176/1 and ST/T00147X/1. The Liverpool Telescope is operated on the island of La Palma by Liverpool John Moores University in the Spanish Observatorio del Roque de los Muchachos of the Instituto de Astrofisica de Canarias. We thank Dr Simone Scaringi and an anonymous reviewer for their careful reading of our original manuscript and their many insightful comments and suggestions that have helped to result in an improved final version.

The NASA/IPAC Extragalactic Database (NED) is funded by the National Aeronautics and Space Administration and operated by the

California Institute of Technology. VSX is the International Variable Star Index data base, operated at AAVSO, Cambridge, MA, USA.

DATA AVAILABILITY

The full list of unclassified GSA sources that our four-class full feature RF model predicted as belonging to the CV class is provided as supplementary material online. A shortened version is shown in Table 8. The LT spectroscopic data for the sources in Table 7 can be acquired from https://telescope.livjm.ac.uk/cgi-bin/lt_search. Both raw and calibrated data files can be obtained by entering the object name in the appropriate field.

REFERENCES

Bellm E. C. et al., 2019, PASP, 131, 018002
Blagorodnova N. et al., 2018, PASP, 130, 035003
Breiman L., 2001, Machine Learning, 45, 5
Cabral J. B., Sánchez B., Ramos F., Gurovich S., Granitto P. M., Vanderplas J., 2018, Astron. Comput., 25, 213
Cao Y., Nugent P. E., Kasliwal M. M., 2016, PASP, 128, 114502
Carrasco-Davis R. et al., 2021, AJ, 162, 231
Chen C., Liaw A., Breiman L., 2004, Using Random Forest to Learn Imbalanced Data. Technical Report 666, Department of Statistics. University of California, Berkeley
Chen T., Guestrin C., 2016, Preprint (arXiv:1603.02754)
Chollet F. et al., 2015, Keras. Available at: https://github.com/fchollet/keras
Copperwheat C. M., Marsh T. R., Dhillon V. S., Littlefair S. P., Hickman R., Gänsicke B. T., Southworth J., 2010, MNRAS, 402, 1824
Cortes C., Vapnik V., 1995, Machine Learning, 20, 273
Cropper M., 1990, Space Sci. Rev., 54, 195
Darnley M. J., Henze M., 2020, Adv. Space Res., 66, 1147
Dieleman S., Willett K. W., Dambre J., 2015, MNRAS, 450, 1441
Downes R. A., Shara M. M., 1993, PASP, 105, 127
Downes R., Webbink R. F., Shara M. M., 1997, PASP, 109, 345
Drake A. J. et al., 2009, ApJ, 696, 870
Fernández A., García S., Galar M., Prati R. C., Krawczyk B., Herrera F., 2018, Learning from Imbalanced Data Sets, 1st edn. Springer, Cham, Switzerland
Förster F. et al., 2021, AJ, 161, 242
Fremling C. et al., 2021, ApJ, 917, L2
Frenay B., Verleysen M., 2014, IEEE Trans. Neural Networks Learning Syst., 25, 845
Freund Y., Schapire R. E., 1997, J. Comput. Syst. Sci., 55, 119
Gabruseva T., Zlobin S., Wang P., 2020, J. Astron. Instrum., 09, 2050005
Gaia Collaboration et al., 2016a, A&A, 595, A1
Gaia Collaboration et al., 2016b, A&A, 595, A2
Goldstein D. A. et al., 2015, AJ, 150, 82
Hellier C., 2001, Cataclysmic Variable Stars – How and Why They Vary. Springer-Verlag, London
Hodgkin S. T. et al., 2021, A&A, 652, A76
Hou W., Luo A. l., Li Y.-B., Qin L., 2020, AJ, 159, 43
Inight K. et al., 2022, MNRAS, 510, 3605
Ivezic Z. et al., 2019, ApJ, 873, 111
Jäger S., Allhorn A., Bießmann F., 2021, Frontiers Big Data, 4, 693674
Jha S. W., Maguire K., Sullivan M., 2019, Nat. Astron., 3, 706
Kato M., Hachisu I., 2012, Bull. Astron. Soc. India, 40, 393
Kato T., Kojiguchi N., 2021, PASJ, 73, 1375
Khan F., Khan K., Singh S., 2018, J. Phys.: Conf. Ser., 1060, 012014
Khazov D. et al., 2016, ApJ, 818, 3
Kochanek C. S. et al., 2017, PASP, 129, 104502
Kruse R., Mostaghim S., Borgelt C., Braune C., Steinbrecher M., 2022, Computational Intelligence: A Methodological Introduction, 3rd edn. Springer-Verlag, New York, p. 53
Kulkarni S. R., 2020, preprint (arXiv:2004.03511)
Law N. M. et al., 2009, PASP, 121, 1395
LeCun Y., Bengio Y., Hinton G., 2015, Nature, 521, 436
Lindegren L. et al., 2021, A&A, 649, A2
Mahabal A. et al., 2019, PASP, 131, 038002
Matheson T. et al., 2021, AJ, 161, 107
Morgan J. S., Kaiser N., Moreau V., Anderson D., Burgett W., 2012, in Stepp L. M., Gilmozzi R., Hall H. J., eds, Proc. SPIE Vol. 8444, Ground-based and Airborne Telescopes IV. SPIE, Bellingham, p. 84440H
Muhammad Ali P., Faraj R., 2014, Data Normalization and Standardization: A Technical Report. Machine Learning Technical Report, Koya University
Neira M., Gómez C., Suárez-Pérez J. F., Gómez D. A., Reyes J. P., Hoyos M. H., Arbeláez P., Forero-Romero J. E., 2020, ApJS, 250, 11
Osaki Y., 1996, PASP, 108, 39
Pala A. F. et al., 2022, MNRAS, 510, 6110
Patterson J., 1994, PASP, 106, 209
Patterson J., Thorstensen J. R., Armstrong E., Henden A. A., Hynes R. I., 2005, PASP, 117, 922
Pedregosa F. et al., 2011, J. Machine Learning Res., 12, 2825
Piascik A. S., Steele I. A., Bates S. D., Mottram C. J., Smith R. J., Barnsley R. M., Bolton B., 2014, in Ramsay S. K., McLean I. S., Takami H., eds, Proc. SPIE Vol. 9147, Ground-based and Airborne Instrumentation for Astronomy V. SPIE, Bellingham, p. 91478H
Rest A. et al., 2014, ApJ, 795, 44
Riello M. et al., 2021, A&A, 649, A3
Rokach L., Maimon O., 2008, Data Mining with Decision Trees: Theory and Applications. World Scientific Press, Singapore
Sánchez-Sáez P. et al., 2021, AJ, 161, 141
Scaringi S., Groot P. J., Knigge C., Lasota J. P., de Martino D., Cavecchi Y., Buckley D. A. H., Camisassa M. E., 2022a, MNRAS, 514, L11
Scaringi S. et al., 2022b, Nature, 604, 447
Simonsen M. et al., 2014, J. Am. Assoc. Var. Star Obser. (JAAVSO), 42, 177
Smartt S. J. et al., 2015, A&A, 579, A40
Soley-Bori M., 2013, Dealing with Missing Data: Key Assumptions and Methods for Applied Analysis. Technical Report No. 4. Boston University, Boston, MA
Solheim J. E., 2010, PASP, 122, 1133
Starrfield S., Iliadis C., Hix W. R., 2016, PASP, 128, 051001
Steele I. A. et al., 2004, in Oschmann J. M., Jr, ed., Proc. SPIE Vol. 5489, Ground-based Telescopes. SPIE, Bellingham, p. 679
Strolger L.-G. et al., 2004, ApJ, 613, 200
Szegedi H., Charles P. A., Meintjes P. J., Odendaal A., 2022, MNRAS, 513, 4682
Szkody P., 1998, in Howell S., Kuulkers E., Woodward C., eds, ASP Conf. Ser. Vol. 137, Wild Stars in the Old West. Astron. Soc. Pac., San Francisco, p. 18
Tachibana Y., Miller A. A., 2018, PASP, 130, 128001
van Roestel J. et al., 2022, MNRAS, 512, 5440
Warner B., 1995, Cataclysmic Variable Stars. Cambridge Univ. Press, Cambridge
Wen Q., Sun L., Yang F., Song X., Gao J., Wang X., Xu H., 2020, Preprint (arXiv:2002.12478)
Wenger M. et al., 2000, A&AS, 143, 9
Zhang Z., 2016, Ann. Translational Medicine, 4, 218
Zwicky F., 1964, Ann. d'Astrophys., 27, 300

SUPPORTING INFORMATION

Supplementary data are available at *MNRAS* online.

RF750_CVs.xlsx

Please note: Oxford University Press is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

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