MONTHLY NOTICES of the Royal Astronomical Society

Automated classification of Chandra X-ray point sources using machine learning methods

Journal:	*Monthly Notices of the Royal Astronomical Society*
Manuscript ID	Draft
Manuscript type:	Main Journal
Date Submitted by the Author:	n/a
Complete List of Authors:	kumaran, shivam; Indian Institute of Space Science and Technology, Department of Earth and Space Sciences Mandal, Samir; Indian Institute of Space Science and Technology, Department of Earth and Space Sciences Bhattacharyya, Sudip; Tata Institute of Fundamental Research, DAA, TIFR Mishra, Deepak ; Indian Institute of Space Science and Technology, Department of Earth and Space Sciences
Keywords:	methods: statistical < Astronomical instrumentation, methods, and techniques, astronomical data bases: miscellaneous < Astronomical Data bases, catalogues < Astronomical Data bases, surveys < Astronomical Data bases, X-rays: general < Resolved and unresolved sources as a function of wavelength

MNRAS 000, 1–11 (2015) Preprint 14 September 2022 Compiled using MNRAS LATEX style file v3.0

Automated classification of *Chandra* X-ray point sources using machine learning methods

Shivam Kumaran[1]★, Samir Mandal[1], Sudip Bhattacharyya[2], Deepak Mishra[3]
[1]*Department of Earth and Space Sciences, Indian Institute of Space Science and Technology, Thiruvananthapuram, 695547, India*
[2]*Department of Astronomy and Astrophysics, Tata Institute of Fundamental Research, Mumbai, 400005, India*
[3]*Department of Avionics, Indian Institute of Space Science and Technology, Thiruvananthapuram, 695547, India*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

A large number of unidentified sources found by astronomical surveys and other observations necessitate the use of an automated classification technique based on machine learning methods. The aim of this paper is to find a suitable automated classifier to identify the point X-ray sources in the *Chandra* Source Catalogue (CSC) 2.0 in the categories of active galactic nuclei (AGN), X-ray emitting stars, young stellar objects (YSOs), high-mass X-ray binaries (HMXBs), low-mass X-ray binaries (LMXBs), ultra luminous X-ray sources (ULXs), cataclysmic variables (CVs), and pulsars. The catalogue consists of $\approx 3,17,000$ sources, out of which we select 2,77,069 point sources based on the quality flags available in CSC 2.0. In order to identify unknown sources of CSC 2.0, we use multi-wavelength features, such as magnitudes in optical/UV bands from *Gaia*-EDR3, *SDSS* and *GALEX*, and magnitudes in IR bands from *2MASS*, *WISE* and *MIPS-Spitzer*, in addition to X-ray features (flux, variability and hardness) from CSC 2.0. We find the Light Gradient Boosted Machine, an advanced decision tree-based machine learning classification algorithm, suitable for our purpose, and achieve 93% precision, 93% recall score and 0.91 Mathew's Correlation coefficient score. With the trained classifier, we identify 54,770 (14,066) sources with more than 3σ (4σ) confidence, out of which there are 32,600 (8,574) AGNs, 16,148 (5,166) stars, 5,184 (208) YSOs, 439 (46) HMXBs, 197 (71) LMXBs, 50 (0) ULXs, 89 (1) CVs, and 63 (0) pulsars. This method can also be useful to reliably identify sources of other catalogues.

Key words: methods: statistical – astronomical data bases: miscellaneous – catalogues – surveys – X-rays: general

1 INTRODUCTION

Huge amount of high-quality data from many astronomical sources are becoming available due to large-scale surveys and an open data access policy. Many of these sources are unidentified. These data have a great potential for discoveries of novel classes of sources, new sources of known classes, new observational phenomena and even perhaps new physics. However, the sheer volume of data necessitates taking an automated approach for source classification. Such an automated classifier can be designed using Machine Learning (ML) methods (Ball & Brunner 2010). ML algorithms are capable of learning patterns in big data and can identify decision boundaries based on the already identified examples. Unlike the manual methods of identifying sources, for example, based on the colour-colour diagram clustering limited to 3-dimension, ML methods can create decision boundary in very high dimension feature space.

In optical/IR astronomy, several works have been done for source identification using machine learning (Krakowski et al. 2016; Tous et al. 2020; Tang et al. 2019; Ćiprijanović et al. 2021). However, in X-ray astronomy, the use of machine learning is relatively less. Farrell et al. (2015) did classification of variable sources in the third *XMM-Newton* Serendipitous Source Catalogue (3XMM) using a Random Forest classifier with timing properties. Zhang et al. (2021) used Random Forest and LogiBoost to classify the sources in *XMM-newton's* 4XMM-DR9 using multiwavelength properties from *GAIA*, *WISE*, and *2MASS*. Classification of X-ray binaries based on whether the compact object is a black hole or a neutron star was done by De Beurs et al. (2022) using *MAXI/GSC* lightcurve. Falocco et al. (2022) used Random Forest and AdaBoost to develop an automated classifier for the identification of AGN and to classify them as Type-I or Type-II AGN further with the data from *XMM-Newton* and SDSS. Tranin et al. (2022) used multiwavelength data to classify sources in *Swift-XRT* and *XMM-Newton* serendipitous Source Catalogues using Naive Bayes classifier. Pattnaik et al. (2021) used the spectrum in the energy range of 5-25 keV from the *Rossi X-ray Timing Explorer (RXTE)* to identify the nature of compact object in the Low Mas X-ray Binary objects using Random Forest classifier.

In this paper, we aim to classify the point X-ray sources in the *Chandra* Source Catalogue (CSC) 2.0, which is the second source catalogue (Evans et al. 2020) of the *Chandra X-ray Observatory*. A unique strength of *Chandra* is its angular resolution ($\leq 1''$), which is substantially better than the FWHM ($6''$; Jansen et al. (2001)) of *XMM-NEWTON* and on-axis angular resolution ($5''$; Aschenbach et al. (1982)) of *ROSAT*. CSC 2.0 contains properties of about 3,17,000 sources from the observations till the end of 2014 and a total sky coverage of 558.65 deg^2. Most of these sources remain unidentified.

To the best of our knowledge, no work of automated classification

★ E-mail: kumaranshivam57@gmail.com

Table 1. Quality flags used to filter sources in CSC 2.0 and their description.

Flag code	Description
pileup_flag	ACIS pile-up fraction exceeds ~10%
sat_src_flag	Saturated source in all observations
conf_flag	Source confused (source and/or background regions in different stacks may overlap)
streak_src_flag	Source located on ACIS CCD read-out streak

of unidentified sources in CSC 2.0 has been published yet. Due to the sub-arcsec spatial resolution and high sensitivity of *Chandra*, the source spatial population density in the CSC 2.0 is the highest among X-ray source catalogues. Thus, CSC 2.0 offers an excellent opportunity for serendipitous discovery of objects of known classes as well as new exotic objects (Martinez Galarza et al. 2019).

In this paper, we discuss the development of an automated classifier based on supervised machine learning algorithms for the point sources in CSC 2.0. The classifier primarily use the features available in CSC 2.0 which are flux in five different bands of *Chandra*'s ACIS instrument and variability properties. In addition to the X-ray features, the source identification can be improved with the use of features available in other wavelengths. We obtain multiwavelength features from *2MASS*, *WISE*, *Gaia*-EDR3, *MIPS-Spitzer*, *SDSS* and *GALEX*. We explore decision tree based supervised machine learning classification algorithms. We find that the Light Gradient Boosted Machine gives the best classification performance.

In the §2, we describe the details of the data, standardizing the data and the method for identifying the training set. In §3, we present various classifier models which we explore, and the methodology for selection and validation of the classifier. In §4, we present the result of the model validation and performance evolution. In §5, we give summary and conclusions.

2 THE DATA

The objective of an machine learning (ML) classification model is to learn the relation between the features of a sample and its class label. In supervised ML methods, this relation is learnt using already labelled samples. For astrophysical objects, the features can be observed properties like magnitudes or flux in various wavelength bands. CSC 2.0 provides a tabulated values of various observed properties. Apart from X-ray features, the source properties in other multiwavelength (MW) bands can be used to further improve the classification.

2.1 X-ray Data

Chandra have two focal plane instruments : Advanced CCD Imaging Spectrometer (*ACIS*) and High Resolution Camera (*HRC*). The *ACIS* instrument observes in broad (b): 0.5-7.0 keV, ultrasoft (u): 0.2-0.5 keV, soft (s): 0.5-1.2 keV, medium (m): 1.2-2.0 keV, and hard (h): 2.0-7.0 keV bands. The *HRC* instrument observes in 0.1-10 keV energy band and is designated as 'W' band. CSC 2.0 was prepared with the observations from *ACIS* and *HRC* till the end of 2014. It contains the information of 3,17,167 sources, out of which 2,96,473 are point sources, selected by the parameter $extent\_flag == 0$ in the catalogue. We further filter the sources based on the quality flags available in CSC 2.0 which are : pileup_flag, sat_src_flag, conf_flag, streak_src_flag (Table 1).

For CSC 2.0, the energy flux in each band is determined using aperture photometry. The source count is derived from an elliptical source region and subtracted by the background count in the surrounding region. To convert the count rate to energy flux, the total count rate is summed up and then scaled by the local ancillary response function. In this work, we use aperture corrected average net-flux in b, u, s, m, and h bands, which are named as b-csc, u-csc, s-csc, m-csc, h-csc (Table 2) respectively.

The variability property in the CSC is calculated using three methods:

(i) Gregory-Loredo variability probability
(ii) Kolmogorov-Smirnov (K-S) test
(iii) Kuiper's test

We use both inter and intra observations variability features in b band as separate parameters. Using the Gregory-Loredo method var_intra_index_b, var_inter_index_b and var_intra_prob_b are calculated. The variability index calculated using Gregory-Loredo variability probability indicates if the source is variable. Kolmogorov-Smirnov (K-S) test gives the feature ks_inrta_prob_b and Kuiper's test gives the feature kp_intra_prob_b. We also use the 1σ standard deviation in the inter observation flux (var_inter_sigma_b) as a feature (see Table 2 for details).

We use Chandra Interactive Analysis of Observations (CIAO-4.14) (Fruscione et al. 2006) to download the data from CSC 2.0 using Astronomical Data Query Language (ADQL)[1].

2.2 Multiwavelength (MW) Data

AllWISE catalogue (Cutri et al. 2021) is the all infrared survey catalogue built by combining the data from *Wide-field Infrared Survey Explorer (WISE)* mission's two all-sky survey projects: *WISE* cryogenic phase (Wright et al. 2010) and the post cryogenic NEOWISE survey (Mainzer et al. 2011). The AllWISE is an all sky infra-red survey at the wavelength bands 3.4 μm, 4.6 μm, 12 μm, and 22 μm named W1, W2, W3 and W3 bands respectively. The AllWISE catalogue contains 747,634,026 sources with limiting sensitivities W1 < 17.1, W2 < 15.7, W3 < 11.5 and W4 < 7.7 magnitude. In our work we use W1, W2, W3 and W4 magnitude from the AllWISE catalogue.

The *Gaia* (Prusti et al. 2016) is an optical telescope launched and operated by European Space Agency. The *Gaia* Early Data Release-3 (*Gaia*-EDR3) (Forveille & Kotak 2021) contains 1,811,709,771 sources and gives magnitudes in three broadband optical passbands, green (G), blue (G_{BP}) and red (G_{RP}) passbands. In this work, we use Gaia-EDR3 G, G_{BP} and G_{RP} band magnitudes. We find the association with *Gaia* using CDS X-match positional cross-match service (Boch et al. 2014) such that the source must be within 3σ positional error of CSC 2.0 and *Gaia* EDR-3 error circle.

The 2 Micron All Sky Survey (2MASS) (Skrutskie et al. 2006) is the survey of entire celestial sphere with a 99.99% sky coverage in the infra-red domain at 1.25μm (J), 1.65μm (H) and 2.16μm (K_s) bands. The survey data ware taken by two identical 1.3 diameter telescope at Arizona and Chile in northern and southern hemispheres respectively. The Survey catalogue contains 470,992,970 point sources.

The Multiband Imaging Photometer (MIPS) onboard *Spitzer* (Rieke et al. 2004; Capak et al. 2013) covers the infrared spectrum in the wavebands of 24 μm, 70 μm, and 160 μm. In this work, we use the 24μm band (bandwidth ~ 5μm) data due to its highest photometric accuracy of $6''$.

[1] https://www.ivoa.net/documents/ADQL/20180112/PR-ADQL-2.1-20180112.html

Table 2. Multi-wavelength features from various catalogues used in this work.

Feature Source	Feature Name	Feature Description
CSC 2.0	gal_l2	Galactic longitude
	gal_b2	Galactic Latitude
	b-csc	Flux in ACIS broad (b) band (0.5-7.0 keV)
	u-csc	Flux in ACIS ultrasoft (u) band (0.2-0.5 keV)
	s-csc	Flux in ACIS soft (s) band (0.5-1.2 keV)
	m-csc	Flux in ACIS medium (m) band (1.2-2.0 keV)
	h-csc	Flux in ACIS hard (h) band (2.0-7.0 keV)
	var_inter_prob_b	Inter-observation variability probability in ACIS b band
	var_inter_sigma_b	Standard deviation in Inter-observation flux variability
	var_inter_index_b	Inter-observation variability index
	var_intra_prob_b	Intra-observation Gregory-Loredo variability probability in b band
	ks_intra_prob	Kolmogorov-Smirnov Intra-observation variability probability b-band
	kp_intra_prob_b	Intra-observation Kupier's test variability probability in b band
	var_intra_index_b	Intra-observation variability index
GAIA-EDR3	G	Gaia Green (G) pass-band magnitude
	Bp	Gaia Blue (G_BP) pass-band magnitude
	Rp	Gaia Red (G_RP) pass-band magnitude
GALEX	FUV	Magnitude in GALEX FUV band
	NUV	Magnitude in GALEX NUV band
SDSS	u-sdss	SDSS u band magnitude
	g-sdss	SDSS g band magnitude
	r-sdss	SDSS r band magnitude
	i-sdss	SDSS i band magnitude
	z-sdss	SDSS z band magnitude
WISE	W1	WISE W1(3.4 micron) band magnitude
	W2	WISE W2(4.6 micron) band magnitude
	W3	WISE W3 (12 micron) band magnitude
	W4	WISE W4 (22 micron) band magnitude
MIPS-Spitzer	24_microns_*MIPS*	Magnitude in 24 micron band of *MIPS* on Spitzer
2MASS	J	J-band (1.235 micron) band magnitude
	H	H-band (1.662 micron) band magnitude
	K_s	Ks-band (2.159 micron) band magnitude
Colour*	B-R	Magnitude in Gaia Bp - magnitude in Gaia Rp
	G-J	Magnitude in Gaia G band - magnitude in 2MASS J band
	G-W2	Magnitude in Gaia G band - magnitude in WISE W2 band
	Bp-H	Magnitude in Gaia G_BP band - magnitude in 2MASS H band
	Bp-W3	Magnitude in Gaia G_BP band - magnitude in WISE W3 band
	Rp-K	Magnitude in Gaia G_RP band - magnitude in 2MASS K band
	J-H	Magnitude in 2MASS H - magnitude in 2MASS H band
	J-W1	Magnitude in 2MASS J - magnitude in WISE W1 band
	W1-W2	Magnitude in WISE W1 - magnitude in WISE W2 band

*The 'colour' features are computed using available magnitude values.

The telescope *Galaxy Evolution Explorer (GALEX)* takes observation in the far ultraviolet (FUV: 1344–1786 Å) and near ultraviolet (NUV: 1771–2831 Å) wavelengths with a resolution of $\sim 4.5''$ (FWHM) and $\sim 6.0''$ (FWHM) respectively (Morrissey et al. 2005).

The Sloan Digital Sky Survey (SDSS; York et al. (2000)) is an extensive photometric and spectroscopic survey with an astrometric accuracy of the order of $0.1''$. The SDSS obtains the data in five optical bands:u, g, r, i and z with the central wavelengths of 3560Å, 4680Å, 6180Å, 7500Å, and 8870Å respectively. The limiting magnitudes of u, g, r, i and z are 21.6, 22.2, 22.2, 21.3, and 20.7 respectively. This work uses the 16th data release of SDSS (SDSS-DR16; Ahumada et al. (2020)). This release includes the data from the previous release combined with the Apache Point Observatory Galactic Evolution Experiment 2 (APOGEE-2) survey and from the Extended Baryon Oscillation Spectroscopic Survey (eBOSS).

We use NASA/IPAC Extragalactic Database (NED) to obtain multi wavelength information for the CSC sources. With the November 2021 release, NED integrated the CSC. With the help of the cross-match algorithm Match Expert (MatchEx; Ogle et al. 2015), 80% of the CSC sources were found to have associations with already existing objects and 20% became new objects in the NED database. We use CSC 2.0 names as the object identifier in NED to obtain the multiwavelength property with identifier based query using `astroquery` package. We obtain multiwavelength data for 2,77,069 sources from NED. However, NED server responded with error messages and the data could not be retrieved for 648 sources. Out of 2,77,069 objects in CSC 2.0, we could find an association for 60% sources in *Gaia-EDR3*, 55% in 2MASS, 43% for *MIPS-Spitzer*, 41% for *WISE*, 24% for *SDSS* and 17% for *GALEX*.

The multiwavelength features used from these catalogues are given in the Table 2. Other than these features, we compute the colours from the magnitude available in different bands and use them as additional features. We use an online multiwavelength visualization tool developed by Yang et al. (2021) to identify the colours that show the best class-wise clustering in a colour-colour diagram.

Table 3. Published catalogues used to identify sources in various classes.

Class	Catalogue source	Catalogue Details	Reference
AGN	VERONCAT	Veron Catalogue of Quasars & AGN, 13th Edition	(Véron-Cetty & Véron 2010)
STAR	SKIFF	Catalogue of Stellar Spectral Classifications	Skiff (2013)
YSO		The Spitzer Space Telescope Survey ...	Megeath et al. (2012)
		The Spitzer/IRAC Candidate YSO Catalogue ...	Kuhn et al. (2021)
HMXB	HEASARC	SMCPSCXMM	Sturm et al. (2013)
		High-Mass X-Ray Binaries Catalogue	Liu et al. (2006)
		INTEGRAL Reference Catalogue	Ebisawa et al. (2003)
		Magellanic Clouds High-Mass X-Ray Binaries Catalogue	Liu et al. (2005)
		IBIS/ISGRI Soft Gamma-Ray Survey Catalogue	Bird et al. (2016)
		INTEGRAL/ISGRI Catalogue of Variable X-Ray Sources	Telezhinsky et al. (2010)
		NGC 3115 Chandra X-Ray Point Source Catalogue	Lin et al. (2015)
		Ritter Low-Mass X-Ray Binaries Catalogue	Ritter & Kolb (2003)
		Low-Mass X-Ray Binaries Catalogue	Liu et al. (2007)
		INTEGRAL Reference Catalogue	Ebisawa et al. (2003)
		XMM-Newton M 31 Survey Catalogue	Pietsch et al. (2005)
		M 31 ... Point Source Catalogue	Hofmann et al. (2013)
		ROSAT All-Sky Survey	Haakonsen & Rutledge (2009)
LMXB	HEASARC	INTEGRAL IBIS Hard X-Ray Survey	Krivonos et al. (2015)
		INTEGRAL IBIS 9-Year Galactic Hard X-Ray Survey Catalog	Krivonos et al. (2012)
		IBIS/ISGRI Soft Gamma-Ray Survey Catalogue	Bird et al. (2016)
		M 31 XMM-Newton ... X-Ray Point Source Catalogue	Shaw Greening et al. (2009)
ULX	ULXRBCAT		Liu & Mirabel (2005)
CV	The Open CV Cat.	The Open Cataclysmic Variable Catalogue	Jackim et al. (2020)
PULSAR	ATNF		Manchester et al. (2005)
	FERMI LAT (4FGL)	Fermi LAT Second Catalogue of Gamma-Ray Pulsars (2PC)	Abdo et al. (2013)

2.3 The Training Set

We aim to classify the point sources in CSC 2.0 in the following classes: Active Galactic Nuclei (AGN), X-ray emitting stars (STAR), Young Stellar Objects (YSO), High Mass X-ray binaries (HMXB), Low Mass X-ray binaries (LMXB), Ultra Luminous X-ray sources (ULX), Cataclysmic Variables (CV) and pulsars. First we prepare a list of already identified sources belonging to these classes. Various published catalogues that we use to identify known sources are given in Table 3.

We cross-match the coordinates of the known sources with all the 2,77,069 sources in our list. We use Astropy[2], which is a Python package to perform cross-matching. We select a cross-match radius of $1''$. In case there are more than one source in the cross-match radius, we consider the source with the least angular separation from the target source. Using this, we identify a total of 7,703 sources of which there are 2395 AGNs, 2790 stars, 1149 YSOs, 748 HMXBs, 143 LMXBs, 211 ULXs, 166 CVs and 101 pulsars. The class-wise percentages of identified sources are given in Table 4. These sources are used to train the supervised machine learning algorithm. In our training set, we have a large fraction of AGNs, stars, and YSOs which comprises of a total of about 80% of the entire training set. The classes LMXB, ULX, CV and pulsar are minorities with populations of only 1-3% of the training set.

We create a data-table with 41 MW features (Table 2) from CSC-2.0, *GAIA*-EDR3, *2MASS*, *SDSS*, *WISE*, *GALEX*, and *MIPS-Spitzer* for 2,77,069 point sources in CSC 2.0. Out of this, we keep a separate data table of 7,703 known sources as the training data set. We attempt to identify the rest (2,69,366) of the sources.

[2] This work made use of Astropy: `http://www.astropy.org`, a community-developed core Python package and an ecosystem of tools and resources for astronomy

3 METHODOLOGY

We prepare the multi-wavelength data for a set of already identified sources. We use this set to train a supervised machine learning model. The model learns the pattern of the features in the training set and identifies the best possible decision boundary in the feature space. For designing the machine learning classification model, we use the python package `Scikit-Learn` (Pedregosa et al. 2011). In `Scikit-Learn`, various ML models are defined as Python class with several options to customize the model. These models also implement general routine functions like 'fit' to train the model and use the trained model to 'perdict' class of a new sample. From `Scikit-Learn`, we test Multi Layer Perceptron, K-Nearest Neighbour, Random Forest and Gradient Boosted Decision Trees. We find that decision tree based models—Random Forest (RF) and Gradient Boosted Decision Tree (GBDT)—perform better than other models. We also explore the Light Gradient Boosted Machine (LightGBM) (Ke et al. 2017), which is an advanced development over GBDT, in this work.

3.1 Classifier Models

3.1.1 Random Forest

Random Forest (RF; Breiman (2001)) is an ensemble of decision tree. Each decision tree is built from a randomly selected bootstrapped sample from the training set. Each tree, thus built is unique in nature and acts as a parallel weak learner. For a given source, each tree votes for it belonging to one of the 8 classes and the fraction of trees out of the entire ensemble voting for a particular class is treated as the class membership probability (CMP) of the given source. For example if 8 out of 10 trees votes for a particular sample to belong to class AGN, then the sample is said to be AGN with a membership probability of 0.8.

Table 4. Number of sources under various classes in the training set.

Class	% of training set	Number of sources
AGN	31%	2395
STAR	36%	2790
YSO	15%	1149
HMXB	10%	748
LMXB	2%	143
ULX	3%	211
CV	2%	166
PULSAR	1%	101
Total training set		**7703**
Unidentified Sources		269366
Total		**277069**

3.1.2 *Gradient Boosted Decision Tree*

Gradient Boosted Decision Trees (GBDT) is an ensemble of weak learners (Friedman 2001). Compared to decision tree, where each tree is built independently, in GBDT the trees are built sequentially based on the error of the previous tree. For each newly constructed tree, a loss is calculated based on the error between the predicted and the true values. In the case of classification algorithm, categorical cross-entropy is used as the loss function and is defined as:

$$Loss = -\sum y_i \times log\hat{y}_i \quad (1)$$

where y_i is a vector of length 8 (number of classes) with 1 for the true value and 0 otherwise. Here, $\hat{y}_i$ is also a vector of length 8 representing the probability of the object to belong to each class. The gradient of this loss function at $(m-1)^{th}$ tree is used to construct a new tree. It is then combined to the previous trees after multiplying it with a weight factor called learning rate η, which varies from 0 to 1. Essentially, each new tree is built to minimize the error from the previous tree. The main advantage of GBDT over RF is that each newly constructed tree uses the loss from the previous tree and thus tries hard to better classify previous incorrectly classified sources. GBDT can also learn more complex decision boundaries compared to RF.

3.1.3 *Light Gradient Boosted Machine*

Light Gradient Boosted Machine (LightGBM) was developed by Ke et al. (2017). LightGBM is an advanced and efficient version of Gradient Boosted algorithms. Compared to GBDT where each feature values are compared at the decision nodes of a tree, LightGBM first discretise the value of the input features and then use these values to construct the decision trees. To make learning more efficient, LightGBM implements two novel techniques, namely, Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). With GOSS, LightGBM downsamples the low-gradient examples and upsamples high-gradient examples which are more difficult to learn. Using EFB, LightGBM bundles the mutually exclusive features (the features that rarely takes zero simultaneously) to reduce the dimension of the feature space. Another major capability of LightGBM is that it can handle missing values. It is very useful as we have a large number of missing values in our dataset. It uses Block Propagation method (Josse et al. 2019). In this method for splitting the nodes of the tree, only the available features are used. Wherever a missing value in a sample is found, it is sent to the side that would minimize the final loss.

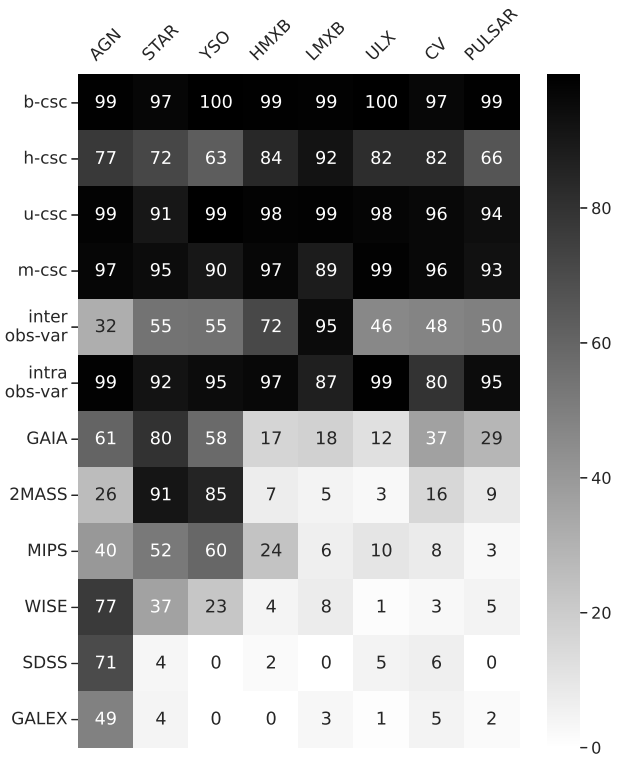

Figure 1. Plot showing the percentage of availability for different features group (see §3.3 for details).

3.2 Data Normalisation

The features value in the data table vary in order of magnitude. Any ML model, in this scenario would tend to give artificially more importance to the feature with high magnitude. Thus it is a general practice to normalise the dataset such that the magnitude variation across feature remains uniform before feeding the data to the model. In our case we normalise the data in such a way that the values lies between 0 and 10 using the following equation,

$$X_{norm} = 10 \times \frac{X - min(X)}{max(X) - min(X)}, \quad (2)$$

where X_{norm} is normalised values of the feature X, $min(X)$ and $max(X)$ is the minimum and the maximum value of the feature X.

3.3 Missing value imputation

In our data-table, we compile features from different multi-wavelength catalogues. Due to the difference in coverage of these catalogues and the differences in limiting sensitivity, objects may not be available in all the catalogues. For example the *SDSS* survey coverage is only limited to northern hemisphere. Similarly, we may have missing values in the data table due to difference in the intrinsic luminosity of the source across different wavelengths. For example, X-ray binaries in quiescent stage have lower luminosity in optical-UV and IR but are prominent in X-rays. In the X-ray domain, based on the variability timescales of the objects, the variability features are not available for some of the objects. The Figure 1 shows the fraction of sources for which the given set of features are available. We can see that 2MASS, *MIPS* and WISE are mostly available only for

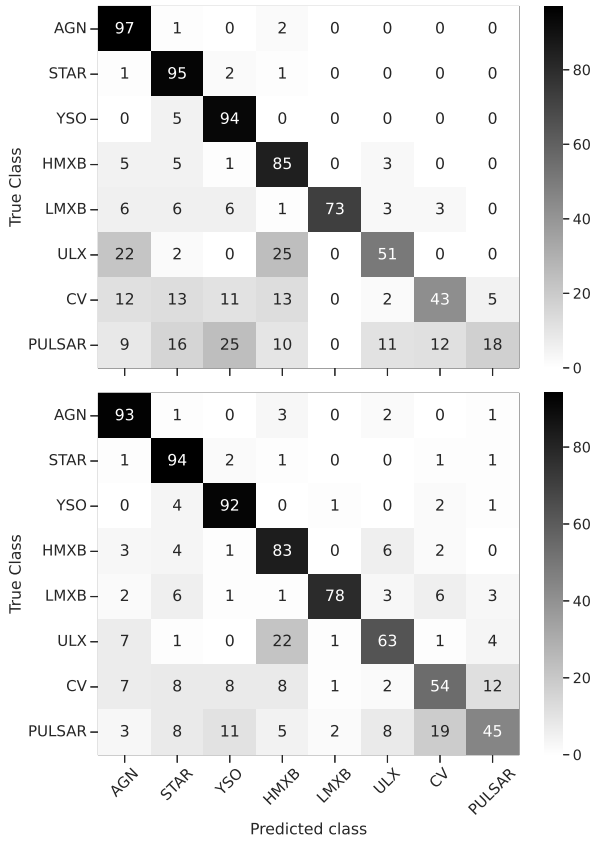

Figure 2. Confusion matrix for Random Forest classifier showing the comparison between no upsampling (top) and upsampling using SMOTE (bottom). The confusion matrix is normalised by the true number of sources available in each class. See §3.4 for details.

AGNs, stars and YSOs. The availability of X-ray variability features are significantly higher for X-ray binaries.

Most of the ML classification models need an input of fixed size and hence are not compatible with missing values. Therefore, these values must be filled prior to training. For RF and GBDT models, we select to impute the missing values using column mode values. However the imputation can create artificial skew in the dataset and make the result less reliable. Moreover, in some cases, the missing values themselves may be important features. For example, X-ray binaries (in quiescent stage) are less likely to be observed in optical wavelengths. Hence to avoid the imputation altogether, we try to design the classifier with Light Gradient Boosted Machine, which can work with missing values in the input feature.

3.4 Class imbalance problem

The Table 4 shows that the numbers of AGNs, stars and YSOs are typically an order of magnitude higher than those of LMXBs, ULXs, CVs and pulsars. Therefore, it is obvious that there is a vast imbalance in the number of training sources, with majority classes being AGN, star, YSO and minority classes are LMXB, ULX, CV and pulsar. Any classifier model can achieve higher accuracy by biasing itself towards the majority class, and thus would fail to perform on the new data.

To tackle the class imbalance problem, we use Synthetic Minority Oversampling Technique (SMOTE; Chawla et al. (2011)). In the feature space, it performs linear interpolations between k-neighbouring points (which represents a source in feature space) and synthetic sources are sampled. Using this technique each class is sampled such that the number of sources in the minority class becomes equal to the same in the most populous class. To keep our result insensitive to the oversampling, we perform SMOTE only on the training set and not on the validation set. SMOTE is used only for the RF and GBDT models. In LightGBM we are working with missing data and SMOTE cannot be performed with missing values. In LightGBM, we use 'class weight' technique which assigns higher weightage to the samples belonging to the minority class in the calculation of loss function (Equation 1) during training. Essentially in the loss function, we make sure that equal contribution comes from each class.

The Figure 2 shows the confusion matrix for the 20-fold cumulative cross validation (discussed in §3.5) without upsampling (top) and with SMOTE upsampling (bottom) for Random Forest model. The true class of the sources is given on the vertical axis whereas the model predicted class is shown on the horizontal axis. The numbers in the matrix elements are in percentage. The diagonal elements in the figure represent the correct classification and the rest are being classified wrongly. We also notice from the top figure that most of the sources belonging to minority classes, mainly CV and pulsar, are getting classified as AGN, STAR and YSO if no upsampling is done. However, the correctly classified pulsars improve from 18% to 45% using SMOTE and $\sim$ 10% improvement happens for other minority classes. In general, the reduced contribution at the lower left corner of the bottom figure as compared to the same in the top figure shows the effectiveness of SMOTE in reducing the bias of the classifier towards the majority class.

Nevertheless any class balancing method is not capable of reproducing a equal class distribution. Both the methods, SMOTE and class-weight are as good as the ability of the available sample to represent the general population of its class.

3.5 Strategy for model performance validation

We compare the performance of the classifiers using a custom version of k-fold cross validation, which we call cumulative k-fold cross validation (CCV).

The flowchart in the Figure 3 shows the cumulative cross validation method. Here, we divide the training set into k-fold, and we train the classifier using $k-1$ folds and keeping aside k^{th} fold as validation set in each iteration. After training we make prediction on the k^{th} fold (represented by the dashed box in the Figure 3). The predictions of classes of these sample are then stored in a prediction table. In the next iteration when a different set of samples are in the validation set, their predictions are then stored in the prediction table. In this manner, after k iterations, in the prediction table we have the predictions for each training source coming only from the iterations when the source was in the validation set. Finally the the elements in the matrix are calculated using this prediction table.

We use precision, recall and F1-score for comparing classifier performance. Precision score is the probability that predicted class is actually the true class for the sample and is defined as:

$$precision_A = \frac{TP}{TP + FP}, \tag{3}$$

where $precision_A$ represents the precision score for class A, TP represents the true positive, i.e., the number predicted as class A actually belonging to class A. Here, FP measures the false positive, i.e., the number of samples for which prediction are class A, while they belong to some other classes.

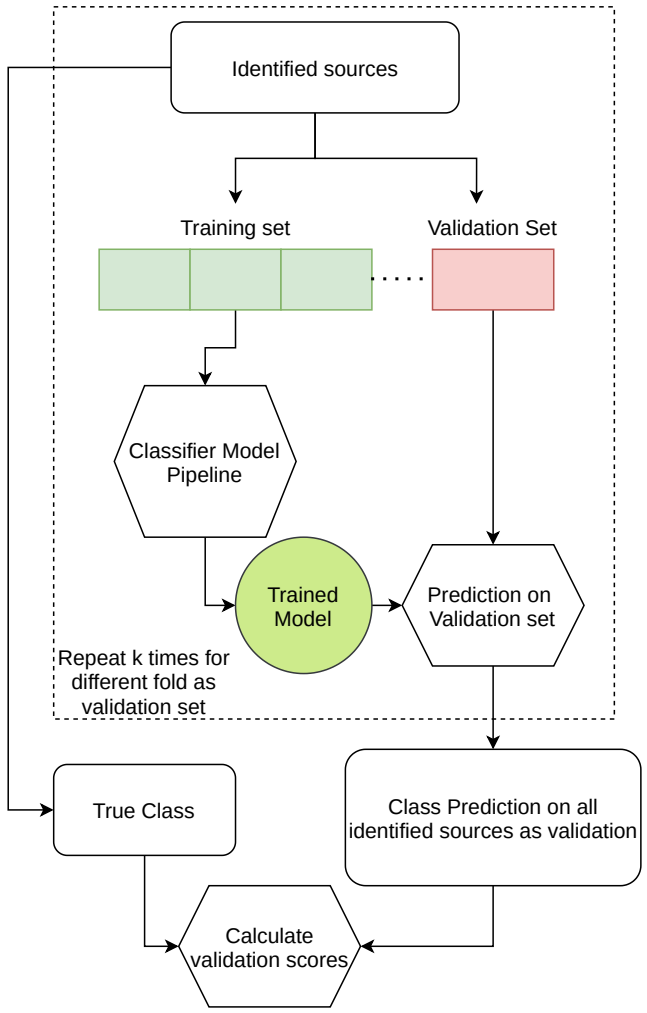

Figure 3. Flowchart showing the cumulative cross validation algorithm. The components inside dashed box represents one fold of validation. The algorithm is discussed in detail in §3.5.

In probabilistic terms, the recall score is the probability of identifying the samples truly belonging to that particular class. The recall score for class A is defined as:

$$recall_A = \frac{TP}{TP + FN}, \tag{4}$$

where FN represents the number of samples belonging to class A but predicted to be in a class other than A.

F1 score is the harmonic mean of precision and recall score:

$$F1_A = 2 \times \frac{precision_A \times recall_A}{precision_A + recall_A}. \tag{5}$$

We also use Mathew's correlation coefficient (MCC), first introduced by Matthews (1975) which is supposed to be a better representation of classification score particularly for imbalanced dataset (Boughorbel et al. 2017). MCC is defined as

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{6}$$

where TN is the number of true negatives. MCC score is comparable to correlation between the output predictions and the true labels whose values vary between +1, 0 and -1. The perfect prediction means +1, 0 means the model is as good as random predictions and -1 shows a complete disagreement between predicted and the true labels. MCC in essence captures the various cases in the confusion matrix.

4 RESULT AND DISCUSSION

We train the classifier models: RF, GBDT and LightGBM, and evaluate the performance using the cumulative k-fold cross validation. We implement the models using `Scikit-Learn` Python library. For the final selected models, we tune the hyperparameters using FLAML (A Fast Library for Automated Machine Learning & Tuning) by Wang et al. (2019). To compare the model performance, using CCV method, we compute the precision, recall and F1 score for each class.

We perform several iterations of the 20-fold CCV, each time starting with a new random seed. These iterations are performed to compute the spread in the scoring metrics till the errors stabilize, which occurred after about 15 iterations. For each model, 15 iterations of CCV are performed and the reported mean and the standard deviations are estimated over these iterations. The precision, recall and F1 score for 20-fold cumulative cross validation are given in Table 5. The performance for majority classes (AGN, star and YSO) with precision, recall and F1 score greater than 92% are significantly higher than ULX, CV and pulsar for which scores are between 40 to 60%. Performance is moderately good for LMXB and HMXB. We observe similar trend across all three models. However, when models are compared, LightGBM performs the best for each class, while RF and GBDT have similarly lower performance.

The LightGBM model performs marginally better than RF for AGN, STAR and YSO but the scores are significantly higher for LMXB, HMXB and ULX. For minority classes, there is a large difference between precision and recall for RF and GBDT model, whereas these two scores are comparable for LightGBM. For example, GBDT tries more aggressively to increase the prediction chances for the minority class and hence improving the recall score by about 7% for ULX, CV and pulsars but this results in a drop in the precision score for the minority classes. Similarly, RF model precision score for pulsars is 35% whereas recall score is 47%. It means that the model is trying hard to predict more pulsars in order to achieve a higher recall score and therefore precision score for pulsar is reduced. This factor is well balanced for LightGBM resulting highest F1 score across all classes.

The Table 6 shows the MCC score and the weighted average of precision, recall and F1 score for RF, GBDT and LightGBM classifier with the weighting factor being the proportion of sample available in each class. The weighted average is taken to account for the class imbalance. LightGBM has the highest F1 score of 93.3% whereas the same for RF and GBDT are about 90%. Once again, it shows that LightGBM classifier performs the best. The MCC score for LightGBM is highest (0.91) and it means that the predictions made by LightGBM is more correlated with the true value compared to other models.

Apart from high precision, recall and F1 score, a good classifier should be capable of predicting the class membership with high confidence. The Figure 4 shows the cumulative histogram of the class membership probabilities (CMP) assigned to the sources in the training set during CCV. The plot shows the histogram for all the three models: RF, GBDT and LightGBM. Any point on the plot shows the number of sources with CMP below the value given on horizontal axis. More steeper the plot towards unity, higher the predicted CMP and more confidant is the model. The figure reveals that the LightGBM is the most confidant classifier model among the

Table 5. Precision, recall and F1 score for different classes for RF, GBDT and LightGBM models. The scores are calculated by 20-fold cumulative cross validation.

Score→	Precision			Recall			F1 score		
class↓ Model→	RF	GBDT	LightGBM	RF	GBDT	LightGBM	RF	GBDT	LightGBM
AGN	96.7±0.1	97.8±0.2	96.8±0.2	93.3±0.2	89.6±0.3	97.6±0.2	95.0±0.1	93.5±0.2	97.2±0.1
STAR	95.6±0.1	97.1±0.2	96.0±0.2	94.1±0.2	91.4±0.2	95.7±0.2	95.0±0.1	94.1±0.1	95.9±0.1
YSO	91.6±0.2	91.2±0.3	92.7±0.3	92.4±0.3	93.7±0.3	95.4±0.3	92.0±0.2	92.4±0.2	94.1±0.2
HMXB	79.2±0.7	83.4±0.8	91.6±0.5	83.4±0.5	87.4±0.7	90.7±0.6	81.2±0.5	85.3±0.5	91.2±0.4
LMXB	84.8±1.3	77.4±2.8	94.8±1.6	80.4±1.1	80.8±0.9	80.9±1.6	82.5±0.8	79.1±1.5	87.2±0.9
ULX	52.4±1.0	47.4±1.1	72.2±1.4	63.5±1.1	75.0±1.1	71.1±1.4	57.4±0.9	57.9±1.0	71.5±1.2
CV	49.1±1.1	42.4±1.0	61.5±1.6	53.8±1.2	60.1±1.6	55.3±1.6	51.4±0.3	49.7±1.1	57.4±1.5
PULSAR	35.8±1.9	28.3±1.2	42.1±1.8	47.1±2.6	55.8±1.8	44.2±1.8	40.7±2.1	37.6±1.4	43.7±2.0

Table 6. MCC score and weighted average precision, recall and F1 score for RF, GBDT and LightGBM models.

Score	Model		
	RF	**GBDT**	**LightGBM**
Precision	90.7 ± 0.1	91.3 ± 0.1	93.2 ± 0.1
Recall	90.0 ± 0.1	89.0 ± 0.2	93.2 ± 0.1
F1 score	90.3 ± 0.1	89.9 ± 0.2	93.2 ± 0.1
MCC	0.87 ± 0.00	0.86 ± 0.00	0.91 ± 0.00

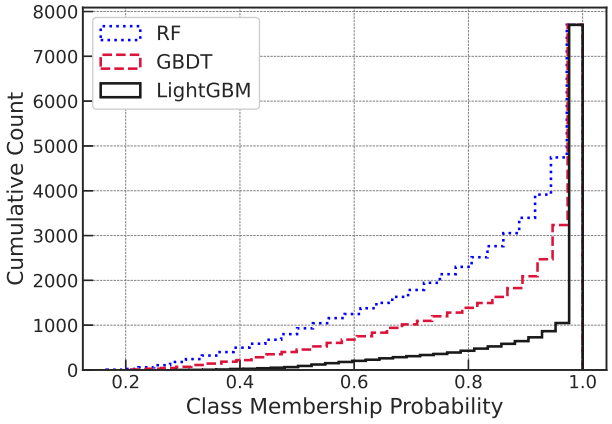

Figure 4. Cumulative histogram of class membership probability for RF, GBDT and LightGBM models, for the training set calculated during cumulative cross validation.

three with only about 500 out of 7703 sources below CMP of 0.8 while RF being the least with more than 2000 sources below CMP of 0.8. Therefore, the cumulative cross validation of the models shows that LightGBM is the best classifier model of choice.

To study the importance of multiwavelength features, apart from the base sample which includes all 41 features mentioned in Table 2, we create another sample keeping only the X-ray features obtained from CSC 2.0. We perform 15 iterations of the 20-fold CCV for the X-ray sample with LightGBM model. Then we compare the model performance with the base sample score to see the class-wise dependency on the MW features which are presented in Table 7 and Figure 5.

The Table 7 shows the mean and standard deviation of the precision, recall and F1 score over 15 times of the 20-fold cumulative cross validation. For each class, the table shows the scores for the

Table 7. Class wise precision, recall and F1 score for CCV using LightGBM model. The validation is done for different set of 'sample type' indicated in the second column: 'all features' where all the 41 features in the Table 2 are used and 'X-ray', the sample with no optical, UV and IR features.

Class	**Sample type**	**Precision**	**Recall**	**F1 Score**
AGN	all-features	96.8±0.2	97.6±0.2	97.2±0.1
	X-ray	91.0±0.2	95.0±0.2	93.0±0.1
STAR	all-features	96.0±0.1	95.7±0.1	95.8±0.1
	X-ray	89.1±0.3	88.2±0.2	88.6±0.2
YSO	all-features	92.7±0.3	95.4±0.3	94.0±0.2
	X-ray	82.9±0.2	89.5±0.5	86.1±0.3
HMXB	all-features	91.6±0.5	90.5±0.5	91.1±0.4
	X-ray	92.0±0.5	89.9±0.4	90.9±0.4
LMXB	all-features	94.7±1.6	80.9±0.7	87.3±0.9
	X-ray	95.0±1.7	82.1±0.5	88.0±0.8
ULX	all-features	72.2±1.4	71.1±1.5	71.6±1.2
	X-ray	61.3±2.0	42.7±2.0	50.3±1.9
CV	all-features	61.5±1.6	55.3±1.7	58.2±1.5
	X-ray	56.0±2.0	44.9±1.8	49.8±1.7
PULSAR	all-features	42.1±1.8	44.2±2.7	43.1±2.0
	X-ray	28.2±1.9	19.0±1.4	22.7±1.4

two cases: one with X-ray features combined with MW features (all-features) and the other with only X-ray features. We can see a clear performance gain for all the classes when the MW features are used combined with X-ray features. The gain is most pronounced for pulsars with an enhancement of F1 score from 22% to 43%. Considering recall score, 44% of the pulsars can be retrieved using all features, however only 19% of the pulsars are classified properly with only X-ray features. For AGN, if all the MW features are simultaneously removed, the precision, recall and F1 score drops by 5%, 2% and 4% respectively. Even with only X-ray features, AGNs have a remarkably high F1 score of 93%. Stars and YSOs too follow a similar trend but the dependency on MW features is higher than AGN, with typically 10% drop in F1 score if MW features are removed. However, HMXB and LMXB are insensitive to optical/UV and IR features and there is hardly any drop in the performance without MW features. This can be attributed to the fact that X-ray binaries are in general faint in optical/UV/IR wavelengths.

For the ULXs, if MW features are removed, F1 score drops about 20% and it is seen that a large percentage of ULXs are classified as AGNs (Figure 5, bottom panel). It results in a slight drop in the F1 score for AGN but it translates to a very high drop in F1 score of ULX due to very small population. For CVs, Optical/UV and IR features are important and F1 score drops by 10% when only X-ray features are used.

Figure 5 shows the confusion matrix for the two cases: one with all

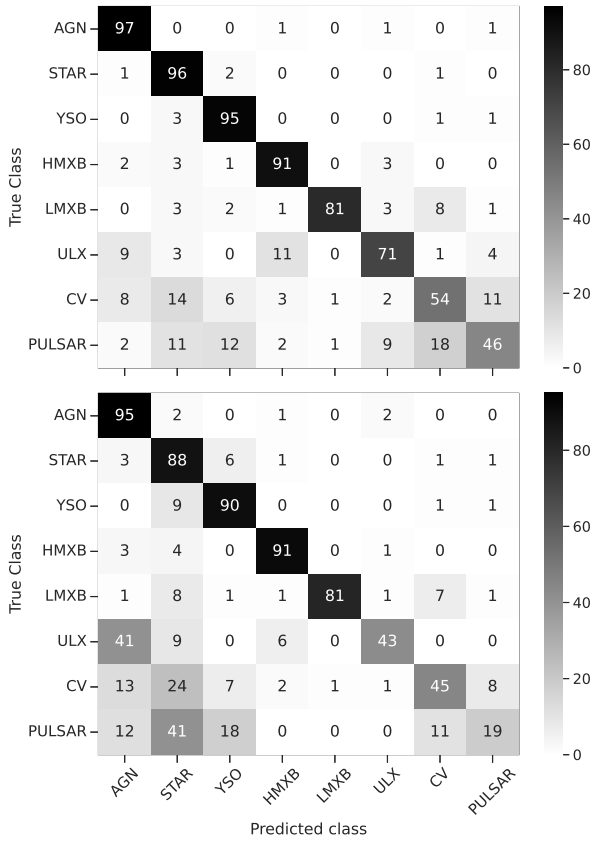

Figure 5. Confusion matrix for the two sample set: all-features (top) and only X-ray feature (bottom). The vertical axis represents the percentage of sources belonging to a class and that being classified into various classes are shown on the horizontal axis.

the features (top) and the other one with only X-ray features (bottom). An element of the matrix shows the percentage of sources which truly belong to the class given on y-axis being classified to a class given on x-axis. The diagonal elements essentially are the recall score for the individual classes. Looking at the ULX row, with MW features only 9% (top figure) of ULXs are identified as AGNs, but the same increases to 41% if only X-ray features are used. Using MW features, the classifier is better able to separate ULXs from AGNs. Similarly, we see without MW features, pulsars are most likely to be identified as stars and the incorrect prediction increases from 11% to 41%. From the confusion matrix, it is evident that the network becomes more biased towards AGN, star and YSO without MW features and the fraction of ULX, CV and pulsar being classified as AGN, star and YSO increases.

From the results discussed above, we select LightGBM as the final classifier with all the 41 features in the Table 2 for training.

The trained LightGBM model is applied to all the 2,69,366 unidentified sources. For each source the trained model gives a probability of belonging to each of the 8 classes. For a given source, the class having the highest probability is assigned to it and the corresponding class probability is called the class membership probability (CMP). Thus we get the class identification and the corresponding CMPs for each of the 2,69,366 sources.

To identify the class-wise confidence of the classifier we dwell deeper into the corresponding predicted probabilities. After the classification is done, we compute the class-wise probability density

Table 8. The number of sources identified in various classes with the LightGBM. The column name 'all' represents the sources classified based on maximum CPM of the class.

Class	Number of sources		
	all	**CMP**$> 3\sigma$	**CMP**$> 4\sigma$
AGN	114642	32600	8574
STAR	63967	16148	5166
YSO	40524	5184	208
HMXB	8321	439	46
LMXB	1688	197	71
ULX	6083	50	0
CV	10999	89	1
PULSAR	23142	63	0
Total	**269366**	**54770**	**14066**

function (PDF) of these CMP. For example, 114642 objects out of 269366 are identified as AGNs (Table 8). Using the CMP of these 114642 objects, we calculate the PDF of AGN's CMP. The probability of getting an object of class A with a CMP = x, lying between the values a and b is given by the area under the curve of the PDF and is given as

$$P(a < x < b) = \int_a^b f_A(x)dx, \tag{7}$$

where the function $f_A(x)$ is the PDF of the class A. It is calculated from the CMP histogram with N bins and counts (n_i) in i^{th} bin using the following equation

$$f_A(x_i) = \frac{n_i}{\sum_0^N n_i \times \Delta x}, \tag{8}$$

where Δx is the bin size. We plot the probability density function in Figure 6 to present the class wise distribution of the CMP. Narrower the peak of the curve towards 1.0, better the prediction confidence. The peak of the PDF is marked by red dashed lines, which is the most probable CMP value for a given class. For AGN, star, YSO, HMXB and LMXB the PDF shows a sharp peak close to 1.0. It means the LightGBM model able to predict most of the AGN, STAR, YSO, HMXB and LMXB with a very high confidence. In fact AGNs are most likely to be identified with the highest probability close to 0.97. However, CMP is ~ 0.5 for ULX, CV and pulsar. To pick the sources which are identified with a very high membership probability, we set a higher threshold to select only those sources above the CMP threshold.

The Table 8 shows the number of newly identified sources in various classes. The number of sources in every class based on the the most probable values of CMP (red line in Figure 6) is given in the second column (all). Similarly, the number of sources above the probability confidence threshold of 3σ (CMP > 0.997) and 4σ (CMP > 0.9999) are shown in the third and fourth column of Table 8 respectively. We identify 54,770 new sources in the existing classes with 3σ confidence and 14,066 new sources with 4σ confidence. This significantly increases the number of sources in various classes.

5 SUMMARY AND CONCLUSIONS

The *Chandra* Source Catalogue CSC 2.0 contains $\approx 3,17,000$ sources, including $\approx 2,77,000$ point sources with a majority of them unidentified. In this work, we implement the decision tree based classifier Light Gradient Boosted Machine to identify the CSC 2.0 objects in the classes of AGN, Star, YSO, HMXB, LMXB, ULX, CV and

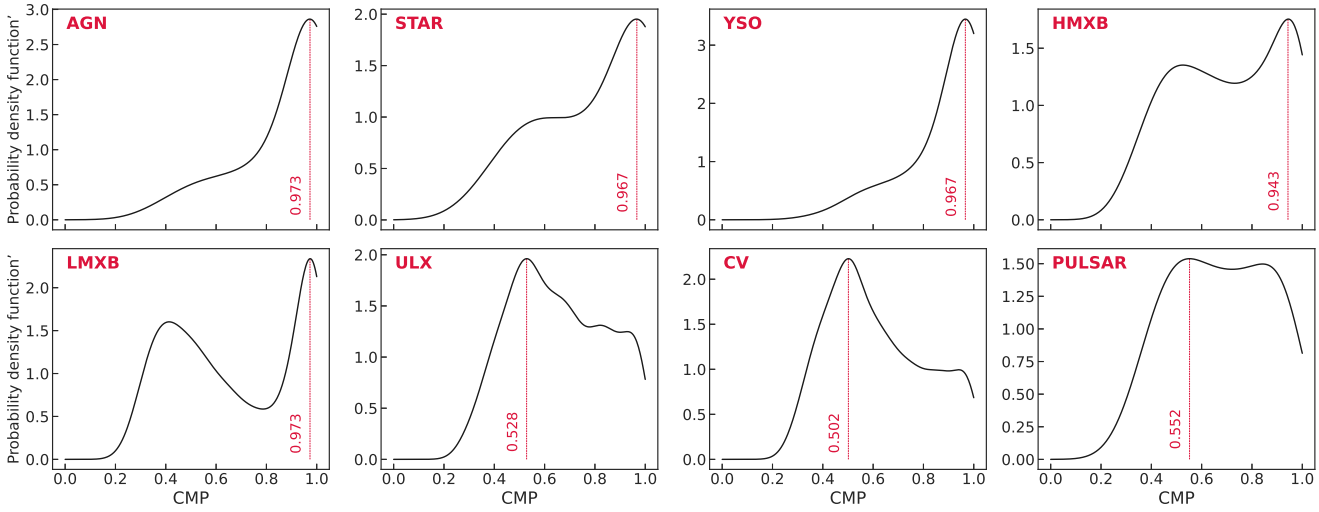

Figure 6. Probability density function (PDF) of the distribution of class membership probability (CMP) of all the unidentified sources predicted using LightGBM model for different classes (marked on the plots). The most probable value of the class membership probability is shown in the plot with vertical dashed line.

pulsar. For the classification, we use X-ray properties from *Chandra*, optical/UV properties from *Gaia*, *SDSS* and *GALEX*, and infrared properties from *2MASS*, *WISE* and *MIPS-Spitzer*. We train the classifier and applied the trained classier to the unidentified sources to estimate class membership probabilities of these sources. We achieve a classification weighted precision score of 93%, recall score of 93%, F1 score of 93% and Mathew's Correlation coefficient of 0.91.

We identify 54,770 new point sources out of which there are 32,600 AGNs, 16,148 stars, 5,184 YSOs, 197 LMXBs, 439 HMXBs, 50 ULXs, 89 CVs and 63 pulsars with a confidence of more than 3σ. Even at a higher confidence (more than 4σ), we get 8,574 AGNs, 5,166 stars, 208 YSOs, 46 HMXBs, 71 LMXBs and 1 CV but not ULXs and pulsars.

The identification of an unknown source as a member of a known class is equivalent to the discovery of a new source of that class. While the aim of this paper is to find a suitable classifier and apply it to CSC 2.0 point sources, in a subsequent paper, we will identify those sources which could be assigned to various classes with high significance values, and thus claim the discovery of many sources of these classes. Finally, we believe that our method can be reliable and promising for other catalogues as well.

ACKNOWLEDGEMENTS

This research has used the *Chandra* Data Archive and the *Chandra* Source Catalogue, and software provided by the *Chandra* X-ray Center (CXC) in the application packages CIAO and Sherpa; 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; data from the European Space Agency (ESA) mission *Gaia* (`https://www.cosmos.esa.int/gaia`), processed by the *Gaia* Data Processing and Analysis Consortium (DPAC,`https://www.cosmos.esa.int/web/gaia/dpac/consortium`); the cross-match service provided by CDS, Strasbourg; Multiwavelength visualisation tool by Yang et al. (2021) to identify the colour-colour properties to be used.

This publication makes use of data products from the Wide-field Infrared Survey Explorer, which is a joint project of the University of California, Los Angeles, and the Jet Propulsion Laboratory/California Institute of Technology, and NEOWISE, which is a project of the Jet Propulsion Laboratory/California Institute of Technology. WISE and NEOWISE are funded by the National Aeronautics and Space Administration.

DATA AVAILABILITY

All the data accumulated/generated in this work will be made available on a public accessible portal which is in development phase. The portal will include the multi wavelength data for the 2,77,069 sources. The portal will also give a facility for the user to perform cone search based on the coordinate. The classification table and the class membership probabilities generated in this work will be integrated in the portal such that the user will be able to filter the search for their class of interest with their selected confidence threshold. The sources identified with a very high confidence will be released as a machine readable data-table in a subsequent paper. The training dataset with the corresponding reference will be shared as a CSV table on the reader's request.

REFERENCES

Abdo A. A., et al., 2013, ApJS, 208, 17
Ahumada R., et al., 2020, ApJS, 249, 3
Aschenbach B., Bräuninger H., Kettenring G., 1982, Advances in Space Research, 2, 251
Ball N. M., Brunner R. J., 2010, International Journal of Modern Physics D, 19, 1049
Bird A. J., et al., 2016, ApJS, 223, 15
Boch T., Pineau F., Derriere S., 2014, CDS xMatch service documentation, `http://cdsxmatch.u-strasbg.fr/xmatch/doc/CDSXMatchDoc.pdf`
Boughorbel S., Jarray F., El-Anbari M., 2017, PloS one, 12, e0177678
Breiman L., 2001, Machine learning, 45, 5
Capak P. L., Teplitz H. I., Brooke T. Y., Laher R., Science Center S., 2013, in American Astronomical Society Meeting Abstracts #221. p. 340.06

Chawla N. V., Bowyer K. W., Hall L. O., Kegelmeyer W. P., 2011, arXiv e-prints, p. arXiv:1106.1813
Ćiprijanović A., et al., 2021, MNRAS, 506, 677
Cutri R. M., et al., 2021, VizieR Online Data Catalog, p. II/328
De Beurs Z. L., Islam N., Gopalan G., Vrtilek S. D., 2022, arXiv e-prints, p. arXiv:2204.00346
Ebisawa K., Bourban G., Bodaghee A., Mowlavi N., 2003, Astronomy & Astrophysics, 411, L59
Evans I. N., et al., 2020, in American Astronomical Society Meeting Abstracts #235. p. 154.05
Falocco S., Carrera F. J., Larsson J., 2022, MNRAS, 510, 161
Farrell S. A., Murphy T., Lo K. K., 2015, ApJ, 813, 28
Forveille T., Kotak R., 2021, A&A, 649, E1
Friedman J. H., 2001, Annals of statistics, pp 1189–1232
Fruscione A., et al., 2006, in Silva D. R., Doxsey R. E., eds, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series Vol. 6270, Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series. p. 62701V, doi:10.1117/12.671760
Haakonsen C. B., Rutledge R. E., 2009, ApJS, 184, 138
Hofmann F., Pietsch W., Henze M., Haberl F., Sturm R., Della Valle M., Hartmann D. H., Hatzidimitriou D., 2013, A&A, 555, A65
Jackim R., Szkody P., Hazelton B., Benson N. C., 2020, Research Notes of the American Astronomical Society, 4, 219
Jansen F., et al., 2001, A&A, 365, L1
Josse J., Prost N., Scornet E., Varoquaux G., 2019, arXiv preprint arXiv:1902.06931, p. 18
Ke G., Meng Q., Finley T., Wang T., Chen W., Ma W., Ye Q., Liu T.-Y., 2017, Advances in neural information processing systems, 30
Krakowski T., Małek K., Bilicki M., Pollo A., Kurcz A., Krupa M., 2016, A&A, 596, A39
Krivonos R., Tsygankov S., Lutovinov A., Revnivtsev M., Churazov E., Sunyaev R., 2012, A&A, 545, A27
Krivonos R., Tsygankov S., Lutovinov A., Revnivtsev M., Churazov E., Sunyaev R., 2015, MNRAS, 448, 3766
Kuhn M. A., de Souza R. S., Krone-Martins A., Castro-Ginard A., Ishida E. E. O., Povich M. S., Hillenbrand L. A., COIN Collaboration 2021, ApJS, 254, 33
Lin D., et al., 2015, ApJ, 808, 19
Liu Q. Z., Mirabel I. F., 2005, A&A, 429, 1125
Liu Q. Z., van Paradijs J., van den Heuvel E. P. J., 2005, A&A, 442, 1135
Liu Q., Van Paradijs J., Van Den Heuvel E., 2006, Astronomy & Astrophysics, 455, 1165
Liu Q. Z., van Paradijs J., van den Heuvel E. P. J., 2007, A&A, 469, 807
Mainzer A., et al., 2011, ApJ, 731, 53
Manchester R. N., Hobbs G. B., Teoh A., Hobbs M., 2005, AJ, 129, 1993
Martinez Galarza J., D'Abrusco R., Civano F., Evans I., 2019, in AAS/High Energy Astrophysics Division. p. 109.29
Matthews B., 1975, Biochimica et Biophysica Acta (BBA) - Protein Structure, 405, 442
Megeath S. T., et al., 2012, AJ, 144, 192
Morrissey P., et al., 2005, ApJ, 619, L7
Ogle P. M., Mazzarella J., Ebert R., Fadda D., Lo T., Terek S., Schmitz M., NED Team 2015, in Taylor A. R., Rosolowsky E., eds, Astronomical Society of the Pacific Conference Series Vol. 495, Astronomical Data Analysis Software an Systems XXIV (ADASS XXIV). p. 25 (arXiv:1503.01184)
Pattnaik R., Sharma K., Alabarta K., Altamirano D., Chakraborty M., Kembhavi A., Méndez M., Orwat-Kapola J. K., 2021, MNRAS, 501, 3457
Pedregosa F., et al., 2011, Journal of Machine Learning Research, 12, 2825
Pietsch W., Freyberg M., Haberl F., 2005, A&A, 434, 483
Prusti T., et al., 2016, A&A, 595, A1
Rieke G. H., et al., 2004, ApJS, 154, 25
Ritter H., Kolb U., 2003, A&A, 404, 301
Shaw Greening L., Barnard R., Kolb U., Tonkin C., Osborne J. P., 2009, A&A, 495, 733
Skiff B. A., 2013, VizieR Online Data Catalog, p. B/mk
Skrutskie M. F., et al., 2006, AJ, 131, 1163
Sturm R., et al., 2013, Astronomy & Astrophysics, 558, A3
Tang H., Scaife A. M. M., Leahy J. P., 2019, MNRAS, 488, 3358
Telezhinsky I., Eckert D., Savchenko V., Neronov A., Produit N., Courvoisier T. J. L., 2010, A&A, 522, A68
Tous J. L., Solanes J. M., Perea J. D., 2020, MNRAS, 495, 4135
Tranin H., Godet O., Webb N., Primorac D., 2022, A&A, 657, A138
Véron-Cetty M.-P., Véron P., 2010, Astronomy & Astrophysics, 518, A10
Wang C., Wu Q., Weimer M., Zhu E., 2019, arXiv e-prints, p. arXiv:1911.04706
Wright E. L., et al., 2010, AJ, 140, 1868
Yang H., Hare J., Volkov I., Kargaltsev O., 2021, Research Notes of the American Astronomical Society, 5, 102
York D. G., et al., 2000, AJ, 120, 1579
Zhang Y., Zhao Y., Wu X.-B., 2021, MNRAS, 503, 5263

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